SORBET: Automated cell-neighborhood analysis of spatial transcriptomics or proteomics for interpretable sample classification via GNN

Shay Shimonov[1†], Joseph M Cunningham[1†], Ronen Talmon[2], Lilach Aizenbud[3,4,5], Shruti J Desai[2], David Rimm[6], Kurt Schalper[6], Harriet Kluger[3,4,5] and Yuval Kluger[1,6*]

[1]Applied Mathematics Program, Yale University, New Haven, CT, USA.
[2]Viterbi Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel.
[3]Department of Medicine, Yale School of Medicine, New Haven, CT, USA.
[4]Yale Cancer Center, Yale School of Medicine, New Haven, CT, USA.
[5]Yale Center for Immuno-Oncology, Yale School of Medicine, New Haven, CT, USA.
[6]Department of Pathology, Yale School of Medicine, New Haven, CT, USA.

*Corresponding author(s). E-mail(s): yuval.kluger@yale.edu;
Contributing authors: shay.shimonov@yale.edu; joseph.cunningham@yale.edu; ronen@ef.technion.ac.il; lilach.aizenbud@yale.edu; shruti.desai@yale.edu; david.rimm@yale.edu; kurt.schalper@yale.edu; harriet.kluger@yale.edu;
†These authors contributed equally to this work.

Abstract

Spatially resolved transcriptomics or proteomics data have the potential to contribute fundamental insights into the mechanisms underlying physiologic and pathological processes. However, analysis of these data capable of relating spatial information, multiplexed markers, and their observed phenotypes remains technically challenging. To analyze these relationships, we developed SORBET, a deep learning framework that leverages recent advances in graph neural networks (GNN). We apply SORBET to predict tissue phenotypes, such as response to immunotherapy, across different disease processes and different technologies including both spatial proteomics and transcriptomics methods. Our results show that SORBET accurately learns biologically meaningful relationships across distinct tissue structures and data acquisition methods. Furthermore, we demonstrate that SORBET facilitates understanding of the spatially-resolved biological mechanisms underlying the inferred phenotypes. In sum, our method facilitates mapping between the rich spatial and marker information acquired from spatial 'omics technologies to emergent biological phenotypes. Moreover, we provide novel techniques for identifying the biological processes that comprise the predicted phenotypes.

Keywords: spatial omics, spatial transcriptomics, graph neural networks, classification, GNN interpretability

1 Introduction

From embryonic patterning to cancer pathology, many biological processes are a consequence of the spatial organization of tissues and their constituent cells. Numerous technologies for spatially resolved, multiplexed molecular profiling of cells ("spatial 'omics") have emerged to interrogate these patterns. In parallel, multiple consortia have been initiated to comprehensively profile the spatial organization underlying physiologic and pathological processes. Computational methods are being developed to query different facets of these data including cell type annotation [1], cell niche clustering [2, 3], and cell-cell interaction [4]. However, few and limited techniques exist to classify tissues profiled with spatial 'omics technologies according to observed phenotypes, such as response to treatment.

The classification of tissues according to observed phenotypes has immediate scientific and clinical implications. For phenotypic prediction, complex biology is frequently simplified to a manageable set of features, such as cell morphology or established clinical biomarkers [5–7]. Computational methods may fully encapsulate the biology represented in high-dimensional datasets; however, current methods represent limited subsets of information, frequently excluding spatial relationships between markers [8], or have low classification accuracy [3, 9]. To address these shortcomings, a few methods have adapted

Graph Neural Networks (GNN) to predict phenotypes [10, 11]. These methods rely on large numbers of patient samples (n = 100s-1000s), extensive computational resources and strong prior information (primarily cell typing), which are not always feasible in clinical settings. These choices limit the applicability of each method across domains with limited prior information or available samples.

Here, we present Spatial 'Omics Reasoning for Binary labEl Tasks (SORBET), a geometric deep learning framework that infers emergent phenotypes, such as response to immunotherapy, from spatially resolved molecular profiling data. We model tissues as a graph of adjacent cells and leverage graph convolutional networks (GCN) [12] to classify tissues. Implicitly, this model learns phenotype-specific cell signatures, which we term cell-niche embeddings (CNE), that synthesize the cell's molecular profile, the molecular profiles of neighboring cells, and the local tissue architecture. We apply SORBET to three datasets comprising tissues derived from non-small cell lung cancer (NSCLC) [13], colorectal cancer (CRC) [14], and metastatic melanoma, respectively (8). These datasets were profiled using three distinct technologies, Imaging Mass Cytometry (IMC), CO-Detection by indEXing (CODEX), and the CosMx Spatial Molecular Imager ('CosMx') platform [14–16], respectively, which profile proteins (IMC, CODEX) and RNA (CosMx). A tissue phenotype is recorded for each dataset including response to immunotherapy (NSCLC, Melanoma) and immune tissue profile (CRC). Classification of tissues according to these phenotypes demonstrates two unique properties that characterize SORBET's functionality: high accuracy across diverse biological contexts and immediate mechanistic interpretability.

2 Results

2.1 SORBET Overview

SORBET is a supervised geometric deep learning framework that learns to accurately classify and analyze spatially resolved single-cell data into two distinct groups. The two groups comprise a binarized outcome derived from a pertinent shared endpoint. In this paper, we primarily apply SORBET to classify responses to immunotherapies for pre-treatment tissues derived from cancer patients (8.8). The input data for SORBET comprise three elements: the binarized endpoint labels for each tissue, the spatial coordinates of the segmented cells within each tissue, and the aggregate expression levels of the molecules (transcriptomics or proteomics) profiled from each cell (8.1). SORBET has three distinct phases: 'Graph Modeling', 'Learning', and 'Reasoning' (8).

During the Graph Modeling phase, tissues are represented as graphs, wherein each cell corresponds to a "node" and physical cell-to-cell contact denotes an "edge" between nodes (**Figure 1a**; 8.2). The biomolecular composition of each cell is taken as the initial representation of its corresponding

node. A graphical model was chosen to jointly capture both cell-intrinsic features and the distinctive structure of each tissue.Rather than analyzing the entire tissue, SORBET is applied to automatically selected subregions or 'subgraphs'. The novel subgraph extraction algorithm applies prior knowledge of the biological question, such as local abundance of a known tumor marker in cancer, to identify relevant sub-regions of the tissue (8.3). Typically, many subgraphs are extracted from each tissue. Note, the algorithm does not guarantee coverage of entire tissues. Incomplete coverage of the tissue is typically associated with poor quality regions (*e.g.* necrotic regions). Extracting and reasoning about subgraphs serves both machine learning and biological purposes. From a machine learning perspective, sampling subgraphs augments the sample size associated with each dataset. From a biological perspective, analyzing tissue sub-regions facilitates localized analysis of heterogeneous cell environments.

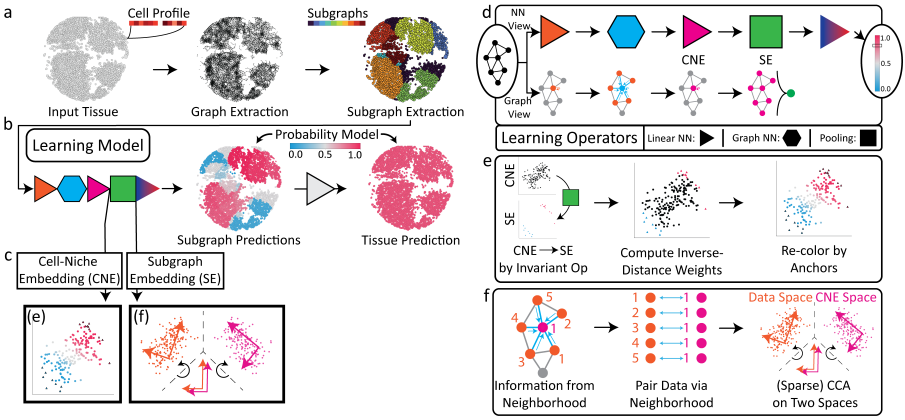

Fig. 1 SORBET Method Overview (a-c) Three SORBET Phases: Graph Modeling, Learning, and Reasoning. **(a)** Graph Modeling phase. Input tissues (left), which are defined by spatially-resolved single cell profiles, are modeled as graphs (middle). From the tissue graph, numerous subgraphs are extracted using an automated subgraph extraction algorithm. **(b)** Learning phase. A graph convolutional network (GCN)-based classifier classifies subgraphs ('Subgraph Prediction', left) and, from these subgraph predictions (middle), a composite tissue prediction is made (right). **(c)** Reasoning phase. The cell-niche embeddings (CNEs) and subgraph embeddings (SEs) are used for tissue prioritization (**(e)**) and marker prioritization (**(f)**). **(d)** Structure of subgraph prediction algorithm. Composition of four neural network operators ('Learning Operators') that are combined to generate a probability. **(e)** Tissue prioritization. The SE space (left, bottom) is generated by a graph invariant operation (*e.g.* `max`) applied to the CNEs (left, top), as in (d). Consequently, the labeled (red / blue) SEs and the unlabeled (black) CNEs occupy the same space (middle). An inverse-distance weighted score (IDWS) is computed for each CNE compared to all SEs. The resulting score indicates the similarity (red / blue spectrum) of each CNE to the relative positions of positively and negatively labeled SEs (red / blue with black outer ring) (right). **(f)** Marker prioritization. Identification of marker programs associated with each phenotype via graph-structured sparse CCA. Input information from graphical neighborhoods (left) is correlated with CNE data (middle) and the resulting canonical correlate weight vectors (right) are used for subsequent analysis.

In the Learning phase, subgraphs are classified using a graph convolutional network (GCN) based architecture (**Fig 1b**; 8.5) [12, 17, 18]. The GCN-based classifier learns by propagating information between cells along the edges of the graph-modeled tissue (blue arrows, **Fig. 1d**; 8.4). As a result, a new embedding is learned for each cell and its niche, which we term cell niche embedding (CNE). Each CNE encapsulates marker data, spatial organization, and phenotypic patterns from the center cell and its surrounding niche. For a specific subgraph, all CNEs are aggregated into a single representation using the `max` operation across the dimensions of the features, which we term subgraph embedding (SE). Consequently, the SEs implicitly reside in the same space as the CNEs. Each SE encapsulates the rich information encoded in the CNEs that constitute the tissue region encompassed by the subgraph. A standard fully-connected neural network classifier maps each SE to the predicted phenotype (*e.g.* response to immunotherapy). To classify whole tissues, we aggregate the predictions from all subgraphs. The collection of SEs and, implicitly, CNEs reside in an embedding space structured to differentiate between phenotypes (**Supplementary Figure A4**, **Supplementary Figure A5**).

In the Reasoning phase, the inferred embeddings are analyzed to understand the biological tissue patterns and mechanisms identified during the Learning phase (**Figure 1c**; 8.6). First, we compute a similarity score between each CNE, which represents local cell-niches, and all SEs, which are labeled according to the observed phenotype (**Figure 1e** 8.6.1). To do so, we compute the inverse-distance weighted sum (IDWS) of each CNE compared to all labeled SEs. This 'Tissue Prioritization' step prioritizes CNEs maximally correlated with either label – relating local tissue patterns to observed phenotypes. Second, we identify marker programs, or groups of markers with shared context, that separate the CNEs and, implicitly, the phenotypes (**Fig. 1f**; 8.6.2). To do so, we apply sparse CCA (sCCA) [19] to compare the input multiplexed data and the CNEs. The sCCA model is applied to each CNE and the cells that comprise its associated niche in pairs (orange (cells) and pink (CNE), **Fig. 1f**). This pairing directly models the graphical propagation of information from surrounding cells to the central cell in the niche (blue arrows). This 'Marker Prioritization' step evaluates the influence of specific markers across neighboring cells at various distances.

2.2 Data Overview

To illustrate SORBET's Learning and Reasoning capacity, we first analyzed spatial transcriptomics data from the CosMx Spatial Molecular Imager platform. This novel dataset comprises 47 patient-derived melanoma samples (**Table B2**; 8.8.3). The CosMx platform targets 960 transcripts. Each patient received one of three immunotherapies (Ipilimumab and Nivolumab combination therapy, Pembrolizumab, or Nivolumab monotherapy; see **Table B1**). Responders (n = 20) were defined as patients demonstrating a partial or complete response to therapy, as determined by comprehensive body scans using standard response criteria [20, 21].

In addition, we utilize two previously reported spatial proteomics datasets from two distinct pathological contexts, Non-Small Cell Lung Cancer (NSCLC) and Colorectal Cancer (CRC) [13, 14] (**Table B2**; 8.8.1, 8.8.2). The NSCLC dataset comprises 29 patient-derived tissues. Protein abundances for 29 markers were characterized using Imaging Mass Cytometry (IMC). Patients received PD-1 / PD-L1 blockade therapy and their response to treatment was recorded (n = 11 responders). The CRC dataset comprises 35 patient-derived tissues. Protein abundances for 49 markers were characterized using the CO-Detection by indEXing (CODEX) technique. These samples were split into phenotypic groups associated with histologic structures, Diffuse Inflammatory Infiltrate (DII, n = 18 patients) and Crohn's Like Reaction (CLR, n = 17 patients). The CLR group, which was associated with positive clinical outcomes, and DII group were defined as the positive and negative phenotypes, respectively. The phenotypic label that defines the CRC task notably differs from the Melanoma and NSCLC datasets – the phenotypic label is defined by the tissue structure rather than response to therapy.

2.3 Performance Results

For each classification task, we consider prediction in two modes: subgraph classification and tissue classification. Subgraph classification comprises prediction of endpoints for tissue subregions, or subgraphs. Subsequently, tissue classification amalgamates predictions from all subgraphs by taking the mean predicted label across all subgraphs. Whereas other choices for amalgamation, such as the maximum predicted label, emphasize single subgraphs, the mean was chosen to emphasize the composite predictions across all subgraphs.

Model performance is assessed using k-fold cross-validation ($k = 5$; **Section** 8.5). The cross-validation procedure is stratified by two factors. First, subgraphs derived from the same tissue are sorted into the same cross-validation fold. This strategy prevents overlaps between data used for training and testing models. Second, positively and negatively labeled samples are balanced across folds. This choice is required due to the small dataset size. For each dataset, cross-validation is repeated ten times. Each repeat is a new random split of the data. The mean and standard deviation of the receiver-operator characteristic (ROC) curve and Area Under the ROC (AUROC) curve are reported across the five repeats.

To the best of our knowledge, supervised learning methods that solely incorporate spatial information and multiplexed 'omics cell profiles are not available. Rather, most existing methods that account for spatial organization rely on additional annotations (*e.g.* cell-typing). For comparison, we consider two methods, CITRUS [22] and Augur [23], that were developed for supervised sample prediction from single cell 'omics without spatial information (8.7). CITRUS (CITRUS-SC) [22], or Cluster Identification, Characterization, and Regression, is an algorithm that first clusters the input data using hierarchical regression, second characterizes all intermediate clusters, and third constructs a regularized logistic regression model on all intermediate clusters (Methods).

Augur (Augur-SC), on the other hand, is a method designed to prioritize cell types based on a given perturbation; similarly, Augur trains a regression model on a set of features that characterize each cluster [23]. In contrast to CITRUS, Augur infers many models across the many possible cell types. Consequently, we developed a third model (Augur(B)-SC) which performs bootstrap aggregation ('bagging') over all trained Augur models. This model was developed to ensure a fair comparison with Augur. This method was primarily developed to prioritize cell types with respect to sample labels. The Augur(B)-SC model is a meta-learning algorithm that combines the individual Augur-SC models to optimize performance.

SORBET displays superior classification accuracy across all three datasets (**Figure 2a-b**). On the Melanoma data set, SORBET achieves an AUROC of 0.83 for subgraph and 0.95 for tissue classifications, outperforming the best performing single-cell subgraph classification method (AUROC = 0.74 by Augur(B)-SC). Even though the same method matches SORBET's tissue classification performance (AUROC = 0.95), SORBET notably excels in the high-confidence domain (FPR < 0.2). This superior performance is also observed in the CRC dataset (AUROC = 0.84 (subgraph), 0.98 (tissue); Augur(B)-SC: 0.76 (subgraph), 0.93 (tissue)) and in the NSCLC classification task (SORBET: AUROC = 0.82 (subgraph), 0.96 (tissue); Augur(B)-SC 0.73 (subgraph), 0.78 (tissue)). These findings affirm the robustness and efficiency of SORBET across diverse classification tasks.

Critically, the choice to classify subgraphs allows for heterogeneous local predictions. To illustrate, we highlight the predictions of SORBET applied to subgraphs on the Melanoma dataset (**Figure 2c**). Here, the model prioritizes multi-cellular sub-regions associated with response (or non-response) to immunotherapy. The chosen examples represent a range of apparent biology from strongly predicted regions (left and right) to graded response (middle-left) to a mixed apparent phenotype (middle-right). The mixed response (middle-right) is associated with non-response to immunotherapy, potentially highlighting tumor heterogeneity that limits therapeutic effect. At a multicellular level, the probability model provides immediate insight into regional biological variation.

2.4 From Predictions To Patterns

The reasoning phase of SORBET comprises two key steps: tissue prioritization and marker prioritization (8.6). In both settings, we introduce novel techniques for interpreting the predictions from each graph neural network. These techniques rely solely on the learned embeddings without the need to re-train additional models or monitor the training process.

In the tissue prioritization step, the model facilitates the identification of important regions across three scales: entire tissues, tissue regions, and individual cell patterns. For tissues and tissue regions, the model prediction provides a well-calibrated measure of the likelihood of the predicted phenotype (8.6). For individual cells, we consider the structure of the shared SE and CNE space.

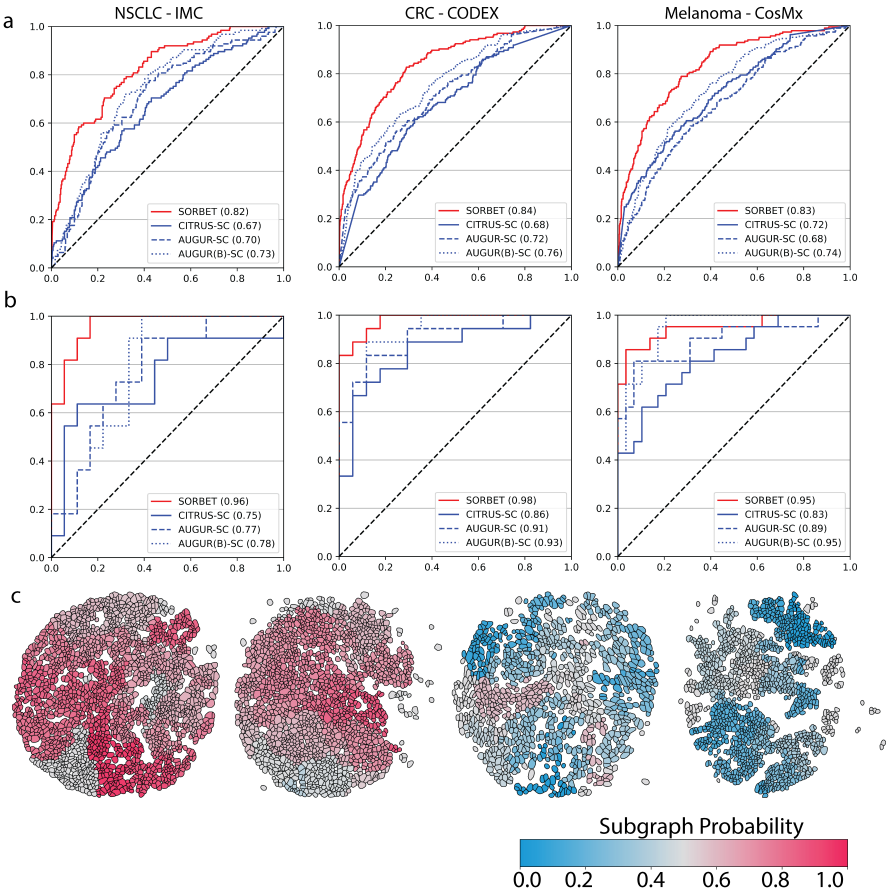

Fig. 2 SORBET classification Accuracy. a-b Columns left to right: Metastatic Melanoma (profiled using CosMx), NSCLC (profiled using IMC), and CRC (profiled using CODEX), respectively. **a.** Subgraph classification task. **b.** Patient classification task. **c.** Examples of subgraph classification, drawn from the Melanoma dataset. Cells not included in any subgraph are set to $p = 0.5$. Left two samples responded to immunotherapy; right two samples did not respond to immunotherapy.

As a supervised model, SORBET optimizes the layout of the SEs to separate between positive and negative labels. Furthermore, the SEs are derived from subsets of the CNEs and, therefore, reside on the same shared embedding space. Intuitively if a CNE is close to positive SEs and far from negative SEs it should be positive. To model this intuition, we adapt an inverse-distance weighted score (IDWS) that accounts for the similarity of each CNE to all SEs (8.6). The inverse-distance weighting assigns a larger weight for SEs that are closer to the CNE and vice-versa. Weights are further modified by a parameter p that controls the locality of the measure. A larger value of p prioritizes nearby points; a smaller value of p averages among a larger number of points.

Biologically, this technique identifies spatial patterns in tissues consistent with the observed phenotypes at the resolution of cell-niches.

In the marker prioritization step, SORBET delineates the key marker programs that define the CNEs using sCCA (8.5). SORBET models each CNE by combining the marker data from the central cell and its nearby cells. First, we compute an sCCA model that correlates the central cell's input features with its corresponding CNE. Next, we correlate the input data from the central cell's immediate neighbors with its corresponding CNE. We progressively expand this process to capture the modeled neighborhoods of the central cell (up to three cells away). This process identifies canonical directions (*i.e.* weight vectors) that maximize the correlation between the input and CNE spaces. Intuitively, canonical directions identified in the input space prioritize marker programs that the model used to separate observed phenotypes. Sparse CCA was chosen to select key subsets of markers most consistent with the output CNE space. The resulting marker programs effectively reflect the relevant information inferred by the model as a function of the distance from the central cell. We apply this technique to all cells and separately for cells associated with either positive or negative labels. By analyzing each potential combination of phenotypes, we prioritize markers and their importance to each category.

2.5 Tissue Prioritization

Here, we analyze each of the datasets using the tissue prioritization method. For each dataset, we plot the two dimensional Uniform Manifold Approximation and Projection (UMAP) [24, 25] applied to the CNEs. The resulting projection is colored according to the computed IDWS ($p = 3$; **Figure 3**). The IDWS of each CNE is mapped to its corresponding central cell in the original tissue.

We consider four tissues contained in the metastatic melanoma dataset, two from patients who responded to therapy and two from patients who did not. Using the IDWS, we identify regions of tissues that most correlate with response (red) and non-response (blue) (**Figure 3a**). Whereas the typical immunofluorescence techniques identify homogenous blocks of tumor and non-tumor regions, SORBET infers heterogeneity among the identified patterns. Visually, tissues labeled as not responding to immunotherapy typically display substantially greater heterogeneity when compared with tissues labeled as responding to immunotherapy. Furthermore, the regions that most strongly correlate with SEs labeled as responding to immunotherapy predominantly overlap with tumor regions in the primary tissue. Nonetheless, SORBET captures a number of phenotypic patterns across all of the tissues that comprise the metastatic melanoma dataset (**Supplementary Figure A1**).

Next, we consider the tissues derived from lung (**Figure 3b**) and colorectal (**Figure 3c**) cancers. The lung cancer cohorts appear to identify supporting structures among the distinct lung architecture associated with response or non-response across tissues. Similarly, the colorectal cancer samples identify regional patterns consistent with the identified phenotypes. These patterns

give rise to the appropriate identification of regions among tissues. Similarly, SORBET captures a wide variety of tissue patterns when many tissues are profiled (**Supplementary Figure A2, A3**).

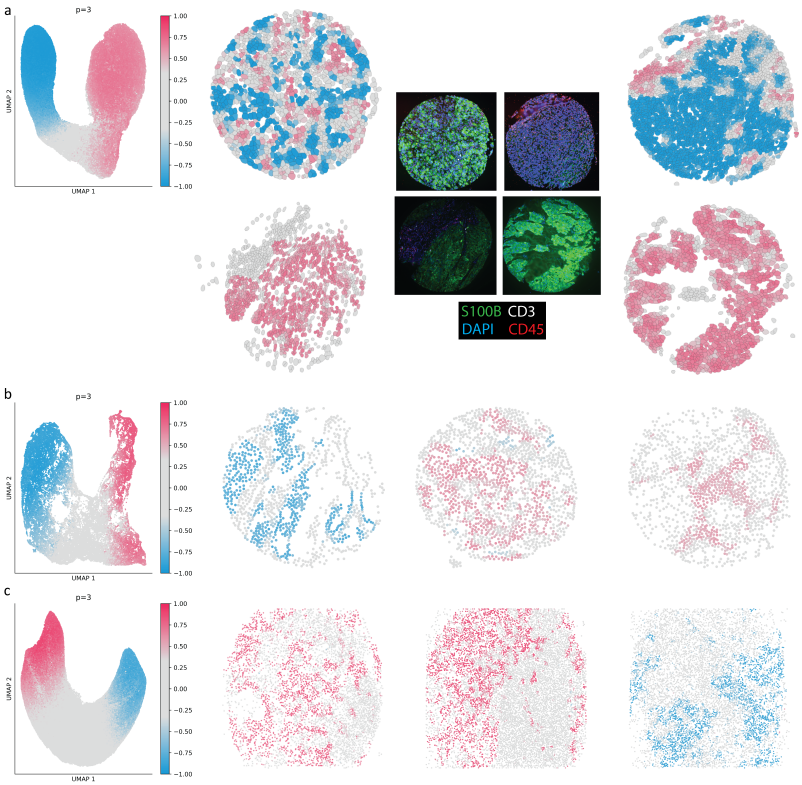

Fig. 3 *Tissue Prioritization* Inverse-distance weighted score (IDWS; $p = 3$) computed for the CNEs across datasets. (Left Column) CNEs are plotted using a two-dimensional UMAP projection with each point colored according to its IDWS. (Right Columns) Tissue examples for each dataset with each cell colored according to its corresponding CNE's IDWS. Colorscale corresponds to observed phenotype (blue - negative, grey - no association, red - positive). Red typically corresponds with response to immunotherapy. **(a)** Melanoma dataset. Tissues represent two non-responders (top) and two responders (bottom). The tissues are accompanied by standard immunohistochemistry images (center). **(b)** NSCLC with three exemplar tissues (non-responder, responder, responder). **(c)** CRC with three exemplar tissues (DII, DII, CLR). Note, CRC was classified according to annotated tissue architecture.

2.6 Marker Prioritization

Next, we analyze each of the datasets using the marker prioritization method. We compute sCCA models, as described, for each of the possible combinations of phenotype groups and distances to the CNE's center cell. The phenotype groups comprise positively-labeled, negatively-labeled, and all-labeled cells.

For each model, we compute the five leading sparse components. We focus on the canonical directions of the input data space, which define the important marker programs. This forms a matrix with each row defining a marker program and each column defining the relative weights of each marker across programs. First, we take the norm over each column (**Figure 4**). The order of column norms prioritizes the markers most consistent with the inferred model. This computation is carried out for each phenotype group. Second, we plot the marker programs from the all-labeled cells phenotype group (**Figure 5**). The plot includes the five leading sparse components for each of the four distances to the CNE's center cell (0,1,2,3).

The magnitude of the computed norms associated with the lung cancer dataset (**Figure 4a**) identifies a number of markers associated with each dataset. Marker prioritization of the NSCLC dataset identifies markers correlated with treatment including *PD1* and DNA1 (a marker of DNA content). Interestingly, this model identifies several immune markers as relevant for predicting response to PD1/PDL1 therapy. In particular, *CD25*, *CD3*, and *CD8* are ranked highly in this dataset, implicating a strong role for adaptive immune cells in this clinical group.

Similarly, re-analysis of the CRC dataset (**Figure 4b**) identifies numerous molecules capturing immunobiology across different scales. For example, three proteins (weight norm ≥ 0.3), *CD2*, *CD45R*α, and *CD45RO*, are important to separate the observed phenotypes and display large differences between weight norms among the CLR and DII groups (weight norm ≥ 0.1). This distinction identifies proteins that capture heterogeneity among the observed data. *CD2*, a protein involved in T-cell activation, displays greater importance (larger weight norm) in capturing the less immune-activated negative label. Similarly, markers for naive (*CD45RO*) and activated (*CD45R*α) T-cells demonstrate meaningful differences among the different regions. Naïve T-cell demonstrate more importance to the biology underlying the positively labeled regions (and the overall prediction) whereas activated T-cells demonstrate more importance for the biology in the negatively labeled regions.

Finally, we analyze the metastatic melanoma data and the most relevant transcripts (**Figure 4c**, **Supplementary Figure A6**). This plot considers the leading 30 positively and negatively extreme markers (with allowed overlap). This analysis identifies numerous markers relevant to immunobiology (*HLA-A/B/C, IGKC*). Perhaps most interestingly, markers associated with cell division (*MZT2A*) and stromal / fibroblastic marker (*VIM*) demonstrate the largest impact on predictions. The leading marker *MZT2A* has previously shown relevance to NSCLC metastasis; however, we were unable to find similar examples implicating its role in melanoma metastasis. These markers provide a rich framework for decomposing the observed melanoma data.

Prioritizing individual markers provides critical insight into parts of the underlying tissue biology. Nonetheless, markers, whether proteins or RNA transcripts, typically function as groups of co-occurring patterns, or programs. As structured, the analysis identifies not just individual markers but sparse

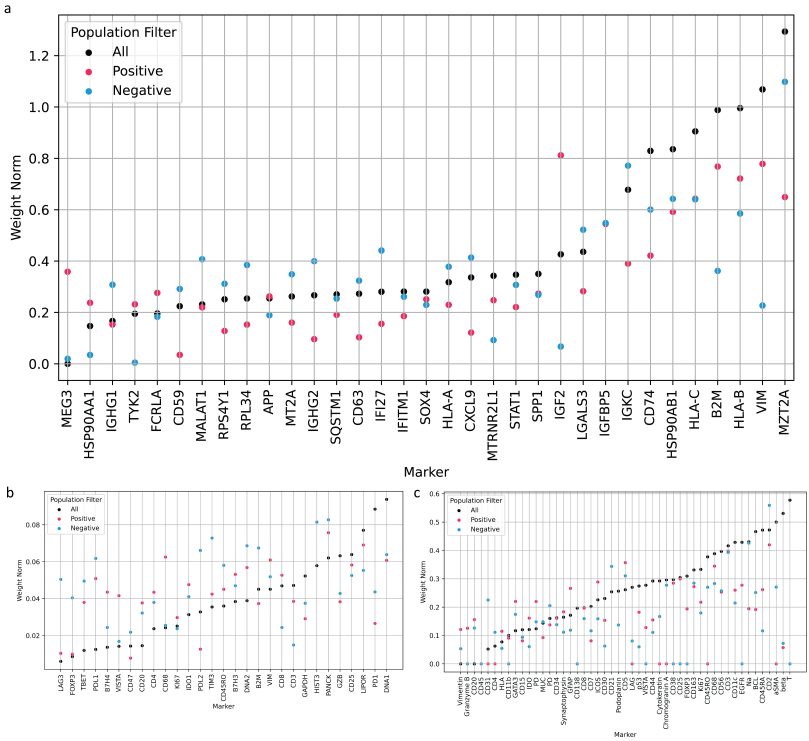

Fig. 4 *Marker Importance Prioritization* Column norm computed over canonical directions for each phenotypic group. Each column represents the relative weights of each marker across marker programs. Phenotypic groups filter the data to include all cells, only positively-labeled cells and only negatively-labeled cells. The computed weight norm (y-axis) is used to sort the markers (x-axis) according to the all cells phenotypic group. **(a)** NSCLC. **(b)** CRC. **(c)** Metastatic melanoma. Metastatic melanoma markers are filtered to the union of the top thirty markers associated with the positively and negatively-labeled cells each.

components representing multiple markers. These sparse components, which represent axes of consistent variation, likely correspond to individual programs (or groups of co-varying programs) that define the observed phenotypes. We plot the leading five sparse components for each of the potential graph distances ($d = 1, 2, 3$ and self-representation $d = 0$) for the different datasets (**Figure 5**). For example, the leading sparse component for adjacent cells in the metastatic melanoma dataset ($d = 1$, index $1 : 0$) identifies *MZT2A* as positively associated with *MALAT1* (as well as *IGF2* and *HSP90AB1*) and as negatively associated with *STAT1*. Putatively, this component may identify an axis (*MZT2A, MALAT1, IFG2, HSP90AB1*) that is associated with positive response; however, this response is negated where growth-factor signaling is predominantly *STAT1*-mediated. Notably, these components are not consistent over larger distances. Biologically, this supports the fact that direct

cell-cell contact differs from paracrine and indirect signaling changes. Technically, this highlights the sophisticated modeling across different scales intrinsic to the reasoning inferred by SORBET.

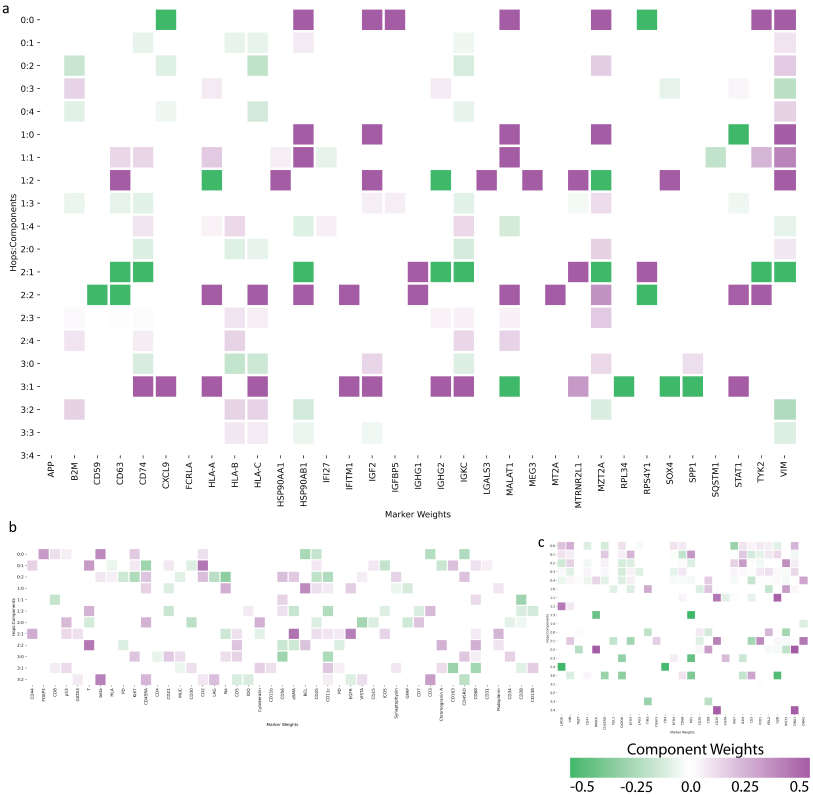

Fig. 5 *Marker Programs* Heatmaps of the canonical directions computed on the input data space for the all-labeled cells phenotypic group. Markers are plotted on the x-axis. The five leading canonical directions are plotted for each of the four distances to the CNE's center cell (0,1,2,3) on the y-axis. The rows are indexed according to distance and canonical direction by the convention '(Distance):(Canonical Direction Index)'. **(a)** NSCLC **(b)** CRC. **(c)** Metastatic melanoma. Metatastic melanoma markers are filtered to the top thirty markers (by column norm).

3 Discussion

SORBET is a geometric deep learning framework that integrates spatial information and 'omics data, enabling effective classification of tissue samples according to observed phenotypes. Furthermore, the structure of SORBET facilitates interpretation of the biological mechanisms driving different phenotypic endpoints. SORBET demonstrates remarkable classification accuracy in various assessments across multiple tasks and spatial 'omics modalities.

The classification accuracy of SORBET, when compared with existing, non-spatial baseline methods, demonstrates the importance of combining spatial and multiplexed molecular information.

Critically, our results are shown on small datasets (n = 29 (NSCLC), 35 (CRC), 47 (Melanoma)) and are computed directly from profiled markers. We introduce automatic subgraph extraction to sample biologically-relevant tissues regions. This approach both captures biological features of the data (*e.g.* tissue heterogeneity) and facilitates improved machine learning performance. Uniquely, automatic subgraph extraction permits deep learning analyses of small datasets characteristic of biomedical applications. As constructed, SORBET applies standard graph neural network methods solely to marker data. These choices allows SORBET to be quickly adapted across distinct biological and technical modalities, including the first prediction of phenotypes from spatial transcriptomics data (the Melanoma dataset). We anticipate that the incorporation of additional tissue information, such as cell morphology, may further improve performance.

The unique architecture of SORBET is tailored to derive structures that correspond to biological entities spanning multiple scales—from individual cells and cell niches to sub-tissue regions and entire tissues. Applied to the three cancer datasets, SORBET identifies mechanisms that are consistent with known tumor biology. SORBET was designed to yield insights relevant for pathologists and biologists. For practitioners, the Tissue Prioritization step highlights regions of interest relevant to the phenotype in question on the original tissues. Similarly, the Marker Prioritization step highlights both individual markers and marker programs that separate the observed phenotype. We believe that the learned cell-niche and sub-region embeddings provide a flexible framework for the exploration of complex biological phenomena with greater depth than explored here.

Although applied to biological analyses here, the Prioritization steps constitute novel methods for analyzing graph neural networks. Existing GNN interpretation frameworks primarily utilize techniques that either require retraining models, which is computationally expensive, or logging intermediate model values, which requires extensive data storage. The proposed methods provide rich interpretation of inferred models with relatively inexpensive computational and data storage costs.

Our work demonstrates the potential SORBET has to offer, especially for pathologists. In clinical settings, non-biochemical techniques, such as morphology and tissue structure, are widely used. They rely on both manual and automatic annotations to identify pathological patterns, such as tumor invasion or necrosis. However, these annotations often fall short in correlating biochemical mechanisms to their consequent tissue patterns. SORBET has the potential to bridge this gap, delivering interpretative insights that seamlessly connect molecules to the biological processes they orchestrate.

4 Data and Code availability

The NSCLC and CRC datasets were previously published. The Metastatic Melanoma dataset will be published pending paper publication. The source code for SORBET is available at https://github.com/KlugerLab/SORBET.

5 Author Contributions

SS, JMC, RT and YK conceived the study. YK provided overall supervision of the study. RT, KS, and HK provided additional supervision. KS provided the NSCLC samples with SJD generating the resulting IMC dataset. DR and HK provided metastatic melanoma samples and clinical data. LA curated clinical annotations and the metastatic melanoma TMA. HK and YK provided the CosMx melanoma dataset. SS and JMC developed the model architecture and interpretation framework, wrote the final code and performed the data analysis. SS, JMC, and YK wrote the manuscript with input from all co-authors.

6 Acknowledgments

HK and YK disclose support for the research of this work from P50CA121974. YK discloses additional support for the research of this work from NIH [R01GM131642, UM1PA051410, U54AG076043, U54AG079759, U01DA053628, and R33DA047037].

7 Competing Interests

The authors declare no competing interests.

8 Methods

8.1 Input Data

SORBET is a framework that classifies binary labeled multiplexed tissues with high precision and explains the biological mechanisms that drive this classification. The framework assumes that the data comprise cells with spatial annotation and assigned multiplexed features. Pre-processing steps to identify cell boundaries, deconvolve overlapping cells, quantify measured markers and associated technical challenges (*e.g.* [26–30]) are not the explicit focus of this work.

Input data are tissues, $\mathcal{T}^{(i)} = (\mathcal{X}^{(i)}, \mathcal{C}^{(i)} | \mathcal{E}^{(i)})$, and binarized labels, $y^{(i)} \in \{0, 1\}$. First, each tissue is defined by cell profiling data, $\mathcal{X}^{(i)} \in \mathbb{R}^{n_i \times M}$, where n_i denotes the number of cells in the tissue, $\mathcal{T}^{(i)}$, and M the number of profiled markers. Second, each tissue is defined by either cell spatial locations, $\mathcal{C}^{(i)} \in \mathbb{R}^{n \times 2}$, or by pre-defined edges, $\mathcal{E}^{(i)} = \{(j, k) \mid j, k \in \{1, 2, \ldots, n\};\ i \neq j\}$. A dataset comprising N tissues is denoted $\{(\mathcal{T}^{(i)}, y^{(i)})\}_{i=1}^{N}$. We call this the tissue classification task.

8.2 Graph Modeling

To translate tissues, $\mathcal{T}^{(i)}$, to a graph, $\mathcal{G}^{(i)}$, we must identify appropriate vertices and edges. Each cell is treated as a vertex, $\mathcal{V}^{(i)} = \{1, 2, \ldots, n\}$, with arbitrary indexing. Datasets may have pre-defined edges $\mathcal{E}^{(i)}$ where the edge $(j, k) \in \mathcal{E}^{(i)}$ if the cells $v_j, v_k \in \mathcal{V}^{(i)}$ are identified as having a physical contact. Otherwise, edges may be inferred from input cell locations $\mathcal{C}^{(i)}$, where an edge $(j, k) \in \mathcal{E}^{(i)}$ if the Euclidean distance $||\mathcal{C}_j^{(i)} - \mathcal{C}_k^{(i)}||_2 \leq \tau$ for some tunable distance threshold τ.

With vertex indexing and edge inference from $\mathcal{T}^{(i)}$, a graph is defined $\mathcal{G}^{(i)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{X}^{(i)})$. The graph construction is independent of downstream analysis. Numerous biologically reasonable graph construction methods might replace this choice.

8.3 Automatic neighborhood extraction

Next, input graphs, $\mathcal{G}^{(i)}$, were automatically processed to extract relevant subgraphs, $SG_j^{(i)} \subseteq \mathcal{G}^{(i)}$. SORBET is applied to subgraphs for two reasons. Biologically, subgraphs provide multiple, localized views of tissues that capture heterogeneity within tissues. Technically, this choice augments available data improving performance and robustness of classifiers. Effectively, this increases the dataset by an order of magnitude.

Formally, a subgraph $S_j^{(i)} = (V_j^{(i)}, E_j^{(i)}, X_j^{(i)}) \subseteq \mathcal{G}^{(i)}$, is defined on a subset of the nodes, $V_j^{(i)} \subseteq \mathcal{V}^{(i)}$ with the edges and marker data appropriately filtered to exclude $v_k \notin V_j^{(i)}$. Consider a function, $f : v_j, \mathcal{G}^{(i)} \mapsto \mathbb{R}$ for $v_j \in \mathcal{V}^{(i)}$, that maps a cell to some real value. The algorithm (**Alg.** 1) identifies 'anchor cells', or cells prioritized by the function f. Around each anchor cell, the median of f is taken on a one hop neighborhood to ensure the cell is not an outlier. Subsequently, a subgraph is formed by taking the k-hop neighborhood around the anchor cell if the neighborhood is of minimum size MS. For this work, we choose $k = 10$ and $MS = 50$. This procedure identifies a set of subgraphs $\mathcal{S}^{(i)} = \left\{S_v | S_v \subseteq \mathcal{G}^{(i)}\right\}$ where S_v denotes the identified subgraph anchored at vertex, v. We associated the graph label, $y^{(i)}$, with each extracted subgraph. Consequently, the sub-tissue classification task comprises the dataset $\bigcup_{i=1}^{N} \left\{(S_v, y^{(i)}) \mid S_v \in \mathcal{S}^{(i)}\right\}$.

The algorithm's output depends substantially on the chosen function, f, and the graph connectivity. Here, we choose a biologically relevant marker (*e.g.* a tumor marker such as *S100B*) and define $f : v_j, \mathcal{G}^{(i)} \mapsto x_j(m)$ where $x_j(m)$ denotes the value of the marker m measured for cell v_j. Note, f is not constrained to adapt a marker value; it may take any relevant value (*e.g.* node degree, centrality).

8.4 SORBET Model Structure

We propose a graph convolutional network (GCN)-based neural network architecture to analyze graphs and subgraphs. The network may be split into four

Algorithm 1 (Automatic Subgraph Extraction):
The function med denotes the median. The function $\mathcal{N}_k$ denotes the k-hop neighborhood around the function argument. Define $\mathcal{G}^{(i)}$ and f as above.

Require: $\mathcal{G}^{(i)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{X}^{(i)})$; f.
Require: Subgraph radius, k, and Minimum subgraph size, MS.
$\mathcal{S}^{(i)} \leftarrow \emptyset$
$T \leftarrow \text{med}_{x \in \mathcal{X}^{(i)}} (f(x))$ (threshold, median of f on all nodes in $\mathcal{X}^{(i)}$).
$Q \leftarrow \left[v_k \mid \forall v_k \in \mathcal{V}^{(i)}, f(v_k) \geq f(v_{k+1})\right]$ (max-queue on $\mathcal{V}^{(i)}$ ordered by f).
$v \leftarrow \text{pop}(Q)$
while $|Q| \neq 0$ **and** $f(v) \geq T$ **do**
    $n_1 \leftarrow \text{med}_{v_j \in \mathcal{N}_1(v)} (f(v_j))$
    if $n_1 >= T$ **then**
        $S_v \leftarrow \mathcal{N}_k(v)$
        if $|S_v| \geq MS$ **then**
            $\mathcal{S}^{(i)} \leftarrow \mathcal{S} \cup \{S_v\}$ (add S_v to the valid subgraphs).
            $Q \leftarrow Q - \mathcal{V}^{(i)}_{S_v}$ (remove the vertices on S_v, $\mathcal{V}^{(i)}_{S_v}$, from Q).
        end if
    end if
    $v \leftarrow \text{pop}(Q)$
end while
return $\mathcal{S}^{(i)}$, the set of all identified subgraphs.

modules: an embedding module, a GCN module, a cell-niche embedding module, and a subgraph prediction module. As above, consider data comprising graphs, $\mathcal{G}^{(i)}$, where $\mathcal{X}^{(i)} \in \mathbb{R}^{n \times M}$.

The embedding module transforms the M-dimensions at each channel into a rich intermediate representation using a fully-connected neural network (FCNN). The FCNN structure applies layer normalization (LN) and the Rectified Linear Unit (ReLU) to each intermediate representation. At each layer, a matrix $W^{(e)}$ of weights is inferred that maps the previous representation, $z_{v,e-1}$, of vertex v to the representation $z_{v,e}$. The initial representation, $z_{v,0}$, is simply the profiled cell data, $z_{v,e=0} = x_v \in \mathcal{X}^{(i)}$. For brevity, we drop the typical superscript, (i), denoting the tissue of origin. Formally, the e-th layer may be written

$$z_{v,e} = \text{ReLU}\left(\text{LN}(W^{(e)} z_{v,e-1})\right)$$

where LN denotes the layer norm $\text{LN}(z) = \frac{z - \mathbb{E}[z]}{\sqrt{\text{Var}(z) + \epsilon}} \cdot \gamma + \beta$ where $\mathbb{E}, \text{Var}$ denote the expectation and variance, γ and β are computed and $\epsilon = 1e-5$ (default) [31].

Next, the GCN module utilizes the Deep GCN layer structure with dense feed-forward connections, as suggested in [17, 18] and implemented in [32]. Briefly, this module infers weights, $W^{(g)}$, that compute the message passing among adjacent vertices. Again, we denote intermediate representations of vertex v at layer g, $z_{v,g}$, and note that $z_{v,g=0} = z_{v,e}$ where e is the final

representation obtained in the embedding module. This layer computes the function,

$$z_{v,e} = \text{ReLU}\left(0, \text{LN}\left(\text{GCN}(z_{v,e-1}, \left\{z_{v_j,e-1} | v_j \in \mathcal{N}_1(v)\right\}\right)\right) \vee z_{v,e-1},$$

where GCN denotes the standard graph convolutional layer, LN the layer norm, and $\vee$ denotes concatenation of two vectors (*i.e.* a residual block connection).

The cell-niche embedding module takes the inferred representations of each node and further compresses the data. This module follows the typical structure of the embedding module. Specifically, it is an FCNN with the same structure (layer normalization and the ReLU activation function). Finally, the subgraph prediction module applies a global non-linearity (global maximum applied along each dimension) across the cell-niche embeddings. Subsequently, predictions are made via a final FCNN with the same structure (layer normalization and the ReLU activation function).

8.5 Learning Phase - SORBET

In this study, we used a stratified 5-fold cross-validation approach to evaluate the performance of our model. Each dataset was split into five groups (approx. 20% of the data) with positively and negatively labeled samples equally split across each fold. For each held-out test fold, appropriate model structure was identified on the remaining folds. Model structure was optimized using the Ray tune parameter optimizer. Model trials were limited to 100 suggested models. Hyperparameters were set to a single set of values across experiments: 50 epochs, 3e−4 learning rate, and 8 batch size. Models were optimized using the Adam optimizer (default aside from learning rate).

This structure differs from traditional approaches to k-fold cross-validation. Typically, three folds might be held-out – two folds to fit model structure and hyperparameters, respectively, and one fold to test. We chose to optimize only model structure due to limited available data. Notably, this choice most likely limits our reported performance. Additional hyperparameter tuning would likely improve test set accuracy.

To accurately assess model performance, we draw 10, random 5-fold cross-validation splits. Models were trained on each of the 50 training sets (ten random splits, five folds each split) and assessed on a held-out test set. Model performance is reported using the receiver operator curve (ROC). Two metrics are returned, the mean (± standard deviation) area under the ROC curve (mAUROC) and a mean (± standard deviation) ROC curve (mROC). Let $g(y, \hat{y}, z)$ denote the estimated true-positive rate according to the ROC curve for labels y and predictions $\hat{y}$ at a false-positive rate z. The mAUROC is computed over the ten random splits, $j = 1, \ldots, 10$,

$$\text{mAUROC}(y, \hat{y}) = \frac{1}{10} \sum_{j=1}^{10} \int_0^1 g(y, \hat{y}, z) dz.$$

Note, the inner integral is simply the typical AUROC; the mean is taken over the ten random splits. The mROC is computed point-wise for 1,000 points uniformly split in the interval $[0, 1]$. At each point, let

$$\bar{g}(y, \hat{y}, z) = \frac{1}{10} \sum_{j=1}^{10} g(y, \hat{y}, z),$$

where $\bar{g}$ denotes the mROC curve. The standard deviations for the mAUROC and mROC are computed in the standard way.

8.6 Reasoning Phase - SORBET

SORBET provides a natural mechanism for understanding the predictions that are inferred during the Learning Phase. In particular, SORBET's Reasoning Phase captures two elements of the problem: the tissue structure and the marker structure. The tissue structure is captured by reasoning about the pooling operation translating between the CNEs and their corresponding SE. The marker structure is captured by adapting a sparse-variance of CCA to reason about the underlying graph structure. We note that both reasoning choices, directly reasoning about pooling operations and using graph-structures sparse CCA to reason about inputs, are general and novel. We believe that these may be of general interest to the machine learning community.

8.6.1 Tissue Prioritization

The relative importance of each tissue structure is defined with relation to the pooling operator employed in SORBET. The pooling operator considers the output from the cell-niche embedding module, which we will denote $z_j^{(i)} \in \mathbb{R}^D$, where D is the output embedding dimension of the cell-niche embedding. The pooling operator, a global max, maps the cell-niches associated with a subgraph to a single representation of each subgraph, $\hat{z^{(i)}}(k) = \max_{j=1,\ldots,n_i} z_j^{(i)}(k)$, where n_i denotes the number of cells and the parenthesis notation, $z(k)$, denotes the k-th component. Notably, each subgraph is summarized by a vector of the same dimension defined on the same support.

The output label prediction is constructed on the space of subgraph summary vectors, $\hat{z^{(i)}}$. Therefore, we compare local cell-niches, defined by the cell-niche embedding module X_i, and compute the relative similarity of these vectors to each summary vector. We compute this similarity using an inverse-distance weighting scheme relative to these summary vectors. Briefly, the inverse-distance score, w, is defined

$$w(z_j^{(i)}) = \frac{\sum_{k=1}^{|\mathcal{S}|} d\left(\hat{z^{(k)}}, z_j^{(i)}\right)^{-p} \mathbb{I}\left(y^{(k)}\right)}{\sum_{k=1}^{|\mathcal{S}|} d\left(\hat{z^{(k)}}, z_j^{(i)}\right)^{-p}},$$

where $\mathcal{S}$ is the set of all subgraphs (as above), $d(\cdot,\cdot)$ is a distance metric, p is a chosen scaling parameter, and $\mathbb{I}$ is an indicator function that maps the binary labels to -1 and 1 (*e.g.* $\mathbb{I}(0) = -1$, $\mathbb{I}(1) = 1$). Note, $w \in [-1, 1]$

Here, we let the distance function be Euclidean distance, p is chosen as reported, and all computations are computed on the leading principal components to 0.01 explained variance ratio.

8.6.2 Marker Prioritization

The relative importance of each marker to inform the model's predictions is similarly defined with respect to the input space of the model. Consider, the GCN-based structure passes information from all nodes within three hops to a central node in the graph. To assess marker importance, it is sufficient to consider the information shared between that central node and the nodes that pass it information. Here, we adapt a sparse-variant of canonical correlation analysis (CCA), $\ell 0$-based Sparse CCA, that decomposes the sparse covariance structure between the inputs and outputs [19].

Let v_i denote the cell under consideration. The CNE associated with v_i, z_i is comprised of information coming from neighboring cells up to three hops away, $N_k(v_i)$ for $k = [1, 2, 3]$, as well as the nodes intrinsic information, x_i. The core problem of computing CCA is to pair information in the input to potentially impacted information in the embedding. In general, a single cell, v_i, has an embedding defined by multiple neighbors. We address this problem by constructing all possible pairs.

Explicitly, we construct four CCA problems, one for each hop and the nodes intrinsic information. For the intrinsic information, Sparse CCA is computed on the input data, X, and the output CNE's, Z, where an immediate pairing exists between the cell's input data x_i and its output embedding z_i. For each k-hop neighborhood, we construct an input matrix, $\tilde{X^{[k]}}$, and an output matrix, $\tilde{Z^{[k]}}$, where for each $v_j \in \mathcal{N}_k(v_i)$, there exists an index l such that $\tilde{X_l^{[k]}} := x_j$ and $\tilde{Z_l^{[k]}} := z_i$. In plain language, we construct matrices where a row in the embedding matrix, or a CNE of a specific cell, correlates with a k-hop neighbor of that cell.

We compute a sparse CCA to identify markers most correlated with each dimension. We adapt an $\ell 0$-based sparse CCA to assess performance [19]. We apply sparse regularization to input dimensions (and not to output dimensions). An appropriate sparsity parameter is chosen using two-fold cross validation. We add an orthogonal regularizer to the cost function to assess additional components. Components are added in a greedy fashion.

8.7 Single-Cell Supervised Prediction Methods

We again note the comparative methods employed in this study, namely CITRUS and Augur. CITRUS, or Cluster Identification, Characterization, and Regression, is an algorithm geared towards the automated discovery of statistically meaningful stratifying biological signatures within extensive single

cell datasets containing multiple known endpoints. It functions in multiple stages starting with unsupervised hierarchical clustering to identify cell clusters within the overall dataset, followed by calculation of biologically pertinent features of these clusters per sample. CITRUS then uses a regularized supervised learning algorithm to identify the populations and features that best predict to which endpoint group a given sample belongs.

Augur, on the other hand, is a method specifically designed to prioritize cell types based on their molecular response to a perturbation in high-dimensional single-cell data. This strategy makes no assumptions about the distribution of features provided as input, and works with various modalities including single-cell RNA-seq, proteomics, epigenomics, and imaging transcriptomics datasets. Augur trains a machine-learning model for each cell type and quantifies separability by assessing how readily the experimental sample labels associated with each cell can be predicted. To control for the uneven representation of cell types in most single-cell datasets, Augur uses repeated sampling of fixed size from each cell-type specific gene expression matrix, providing a quantitative basis for cell type prioritization. To enhance computational efficiency, Augur incorporates two feature selection steps to minimize the number of analytes provided to the classifier as input.

Finally, we combine many Augur predictors using the standard machine learning "meta-learning" technique, stacking. Stacking combines predictors by finding weights that optimally combine the underlying predictors.

8.8 Datasets

Here, we describe three datasets and their associated classification tasks drawn from different biological settings (Non-small cell lung cancer, Colorectal cancer, Metastatic melanoma), three different technologies (Imaging Mass Cytometry, Co-detection by indexing (CODEX), and CosMx Spatial Molecular Imager (SMI)) and two different biomolecular types (proteomics, transcriptomics). Sub-sections are annotated by biological setting, technology and biomolecular type.

8.8.1 Non-small Cell Lung Cancer / Imaging Mass Cytometry Spatial Proteomics Classification

The first classification task, classification of patient-derived non-small cell lung cancer (NSCLC) tissues, comprises a previously described dataset [13]. Protein abundances for 33 total markers were quantified using imaging mass cytometry (IMC). The data comprise patients (n = 29) treated with PD-1 / PD-L1 immunotherapy. Patients were categorized according to response to immunotherapy as measured by full-body scans (n = 11 responders). The training task comprises analyzing

Raw samples were processed as follows. Four markers, Xe-131, Xe-132, Hg-200, and 190BCKG, which represent standard background markers, were excluded. Consequently, 29 markers were used in the NSCLC analysis task.

Anchor cells were identified using the pan-cytokeratin (PanCK) tumor marker and subgraphs were extracted using the described subgraph extraction algorithm. For these data, the cell contacts were pre-annotated. Per patient, an average of 13.4 (STD: 5.6) subgraphs were extracted. Ten-fold cross-validation was performed, which generated an average split of 23, 3, and 3 for train, test and validation, respectively.

For comparison, we consider four different methods applied to the same task. First, we apply a method previously developed for analysis of patient-derived IMC tissues derived from breast cancer patients [13]. Briefly, this technique classifies tissues in two steps. First, a random walk is computed on the graph extracted from the underlying data until a steady-state distribution is reached. Second, the steady state distribution is embedded and the embedding space is subsequently used to classify the original data. Then, we considered the three standard baselines applied to all of the classification tasks as described.

8.8.2 Colorectal Cancer / CODEX Spatial Proteomics Classification

The second classification task, classification of patient-derived colorectal cancer (CRC) samples, is derived from the work of Schurch, *et. al.* [3]. We do not apply any additional data processing to the raw data extracted from that work. Briefly, Schurch, *et. al.* extracted four tissue samples from FFPE samples for each patient (n = 35). Patients were categorized as displaying a Crohn's like reaction (CLR) if two of the four tissue samples displayed tertiary lymphoid structures (TLS) and two of the four tissue samples displayed diffuse inflammatory infiltrates (DII) by hematoxylin and eosin (H&E)-stained tissue samples. Patients were categorized as displaying a DII reaction if all four tissue samples displayed DII. The CLR reaction was associated with improved survival rates when compared with the DII population. Subsequently, 56 proteins were quantified in all tissues using the CODEX technology. CODEX iteratively quantifies protein abundance via annealing of DNA-conjugated antibodies and (complementary) DNA-conjugate fluourescent probes. Raw images were processed using the CODEX Toolkit [14]. Cells were segmented based on the DRAQ5 nuclear stain using the CODEX Toolkit segmenter.

The processing conducted by Schurch, *et. al.* generated 35 patient samples with four regions for each patient separated into a Crohn's like reaction (CLR, n = 17) and DII (n = 18). Anchor cells were identified using the pan-cytokeratin (PanCK) tumor marker and subgraphs were extracted using the described subgraph extraction algorithm. Cells contacts were inferred using an estimated distance threshold, τ. For a sample with n nodes, τ is defined as the $\frac{10}{n}$-th percentile distance from the sorted list of all cell-cell distances. Cell counts, c_i are log-transformed, $c_i \rightarrow \log(c_i + 1\text{e}{-3})$, and z-normalized, as in the original paper. Per patient, an average of 35.9 (STD: 7.3) subgraphs were extracted. The model was applied to classify tissues as displaying a CLR or a

DII-reaction. Ten-fold cross validation was performed, which generated average splits of 28, 4, and 3 for training, validation, and test sets, respectively. Each patient was assigned to a single test set.

For comparison, we consider four different methods applied to the same task. First, we implement the classification approach proposed in the initial paper in Python using the `glmnet` package. The cells are grouped into distinct cell types. Marker counts are summed for each cell type. Counts are either extracted from the complete tissue (cell-type, CT) or per annotated region (cell-neighborhood cell-type, CN-CT) in the underlying tissue. Counts, c_i, are log-transformed, $c_i \rightarrow \log(c_i + 1\text{e}{-3})$, and z-normalized for each marker. Normalized log counts are subsequently used to train an L1-regularized logistic regression to predict the same classes as above. The original paper implements repeated hold out for model validation and testing; we reimplement the same model and use 5-fold cross validation for comparison with our method. Then, we considered the three standard baselines applied to all of the classification tasks, as described.

8.8.3 Metastatic Melanoma / CosMx Spatial Transcriptomics Classification

The third classification task, classification of response to immunotherapy from patient-derived melanoma samples, comprises a novel dataset. Initially, a combined 50 formalin-fixed paraffin embedded (FFPE) tumor samples, which were part of a larger tumor microarray (TMA), were drawn from 49 individuals (one patient was sampled twice). Two samples were excluded for poor data quality and one replicated sample was excluded (see below). Consequently, the dataset comprises 47 samples derived from 47 patients. Patients were treated with one of three immunotherapies: ipililimumab-nivolumab (n = 21), pembrolizumab (n = 16), and nivolumab monotherapy (n = 10). Responders (n = 20) were defined as individuals displaying partial or complete response as measured by full-body scan. Median progression-free response times were 426.5 days (336, 796; 95% CI by 1,000 Bootstrap Samples) compared with 81.0 days (66, 136; 95% CI) for non-responders (**Supplementary Figure A7a**). Data characteristics are described in Supplementary Table 1.

Samples were sequenced using the CosMx Spatial Molecular Imager ("CosMx") platform. CosMx sequences FFPE tissue samples via cyclic fluorescence *in situ* hybridization (FISH) techniques that iteratively identify transcripts and their location. For this dataset, the platform sequenced 960 targeted transcripts. In addition, the platform is capable of identifying a limited number of protein analytes via fluourescent immunohistochemistry; here, four cellular markers, *S100B*, *CD3*, *CD45*, and *DAPI*, were measured. The samples were sequenced over two sequencing runs (n = 26, 24, respectively). Counts per cell and negative probes identified per cell are similar between the two runs (**Supplementary Figure A7c-d**). No batch effect corrections were applied to the data.

Raw samples were processed as follows (**Supplementary Figure A7b**). First, sequencing samples were filtered to exclude cells overlapping from adjacent tumor samples in the TMA. Second, cells with fewer than ten transcripts associated with the cell were excluded. Otherwise, no additional tumor processing was conducted. Tumors were modeled as graphs via consideration of inter-cellular distance. The mean radius, r, was computed for all cells in the graph and an edge was assumed between pairs of cells within $3r$ of each other. As described, two data samples were excluded for poor data quality and one sample was excluded as a (biological) replicate. The first excluded sample displayed poor sequencing depth. The second excluded sample displayed poor inter-cellular connectivity. For the biological replicate, one of the two samples was arbitrarily chosen and dropped. The samples exhibit similar sequencing quality and tissue structure. See **Supplementary Figure A7c-e** for an illustration of the excluded tissue samples.

Anchor cells were identified using the mean *S100B* protein abundance as a tumor marker and subgraphs were extracted using the described subgraph extraction algorithm. Cell counts, c_i are log-transformed, $c_i \rightarrow \log(c_i + 1\text{e}{-3})$ and z-normalized. Per patient, an average of 9.1 (STD: 2.1) subgraphs were extracted. The model was applied to classify tissues as likely or unlikely to respond to immunotherapy. Five-fold cross validation was performed, which generated an average of 28-29, 9-10, 9-10 tissue samples for training, validation, and test sets, respectively. Each patient was assigned to a single test set. For comparison, we considered the three standard baselines applied to all of the classification tasks, as described.

References

[1] Brbić, M., Cao, K., Hickey, J.W., Tan, Y., Snyder, M.P., Nolan, G.P., Leskovec, J.: Annotation of spatially resolved single-cell data with STELLAR. Nature Methods **19**(11), 1411–1418 (2022). https://doi.org/10.1038/s41592-022-01651-8

[2] Fang, Z., Ford, A.J., Hu, T., Zhang, N., Mantalaris, A., Coskun, A.F.: Subcellular spatially resolved gene neighborhood networks in single cells. Cell Reports Methods **3**(5), 100476 (2023). https://doi.org/10.1016/j.crmeth.2023.100476

[3] Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., Kinoshita, S., Samusik, N., Goltsev, Y., Nolan, G.P.: Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell **182**(5), 1341–135919 (2020). https://doi.org/10.1016/j.cell.2020.07.005

[4] Fischer, D.S., Schaar, A.C., Theis, F.J.: Modeling intercellular communication in tissues using spatial graphs of cells. Nature Biotechnology **41**(3), 332–336 (2023). https://doi.org/10.1038/s41587-022-01467-z

[5] Frank, R., Hargreaves, R.: Clinical biomarkers in drug discovery and development. Nature Reviews Drug Discovery **2**(7), 566–580 (2003). https://doi.org/10.1038/nrd1130

[6] Belter, B., Haase-Kohn, C., Pietzsch, J.: Biomarkers in Malignant Melanoma: Recent Trends and Critical Perspective. In: Ward, W.H., Farma, J.M. (eds.) Cutaneous Melanoma: Etiology and Therapy. Codon Publications, Brisbane (AU) (2017). https://doi.org/10.15586/codon.cutaneousmelanoma.2017.ch3

[7] Kumar, V., Abbas, A.K., Aster, J.C.: Robbins Basic Pathology, 10th edn. Elsevier - Health Sciences Division, Philadelphia, PA (2017)

[8] Allen, C., Chang, Y., Ma, Q., Chung, D.: Maple: a hybrid framework for multi-sample spatial transcriptomics data. bioRxiv (2022). https://doi.org/10.1101/2022.02.28.482296

[9] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (UK) (2000). https://doi.org/10.1017/CBO9780511801389

[10] Wu, Z., Trevino, A.E., Wu, E., Swanson, K., Kim, H.J., D'Angio, H.B., Preska, R., Charville, G.W., Dalerba, P.D., Egloff, A.M., Uppaluri, R.,

Duvvuri, U., Mayer, A.T., Zou, J.: Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. Nature Biomedical Engineering **6**(12), 1435–1448 (2022). https://doi.org/10.1038/s41551-022-00951-w

[11] Wang, Y., Wang, Y.G., Hu, C., Li, M., Fan, Y., Otter, N., Sam, I., Gou, H., Hu, Y., Kwok, T., Zalcberg, J., Boussioutas, A., Daly, R.J., Montúfar, G., Liò, P., Xu, D., Webb, G.I., Song, J.: Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. npj Precision Oncology **6**(1), 1–12 (2022). https://doi.org/10.1038/s41698-022-00285-5

[12] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=SJU4ayYgl

[13] Lin, Y.-W.E., Shnitzer, T., Talmon, R., Villarroel-Espindola, F., Desai, S., Schalper, K., Kluger, Y.: Graph of graphs analysis for multiplexed data with application to imaging mass cytometry. PLoS computational biology **17**(3), 1008741 (2021). https://doi.org/10.1371/journal.pcbi.1008741

[14] Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., Nolan, G.P.: Deep profiling of mouse splenic architecture with codex multiplexed imaging. Cell **174**(4), 968–98115 (2018). https://doi.org/10.1016/j.cell.2018.07.010

[15] Chang, Q., Ornatsky, O.I., Siddiqui, I., Loboda, A., Baranov, V.I., Hedley, D.W.: Imaging mass cytometry. Cytometry Part A **91**(2), 160–169 (2017). https://doi.org/10.1002/cyto.a.23053

[16] He, S., Bhatt, R., Brown, C., Brown, E.A., Buhr, D.L., Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R.G., Geiss, G., Gregory, M.T., Hoang, M.L., Khafizov, R., Killingbeck, E.E., Kim, D., Kim, T.K., Kim, Y., Klock, A., Korukonda, M., Kutchma, A., Lewis, Z.R., Liang, Y., Nelson, J.S., Ong, G.T., Perillo, E.P., Phan, J.C., Phan-Everson, T., Piazza, E., Rane, T., Reitz, Z., Rhodes, M., Rosenbloom, A., Ross, D., Sato, H., Wardhani, A.W., Williams-Wietzikoski, C.A., Wu, L., Beechem, J.M.: High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nature Biotechnology **40**(12), 1794–1806 (2022). https://doi.org/10.1038/s41587-022-01483-z

[17] Li, G., Müller, M., Thabet, A., Ghanem, B.: DeepGCNs: Can GCNs Go as Deep as CNNs? arXiv. arXiv:1904.03751 [cs] (2019). https://doi.org/10.48550/arXiv.1904.03751. http://arxiv.org/abs/1904.03751

[18] Li, G., Xiong, C., Thabet, A., Ghanem, B.: DeeperGCN: All You Need to Train Deeper GCNs. arXiv (2020). https://doi.org/10.48550/arXiv.2006.07739. http://arxiv.org/abs/2006.07739

[19] Lindenbaum, O., Salhov, M., Averbuch, A., Kluger, Y.: L0-sparse canonical correlation analysis. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=KntaNRo6R48

[20] Kluger, H.M., Zito, C.R., Turcu, G., Baine, M.K., Zhang, H., Adeniran, A., Sznol, M., Rimm, D.L., Kluger, Y., Chen, L., Cohen, J.V., Jilaveanu, L.B.: PD-L1 Studies Across Tumor Types, Its Differential Expression and Predictive Value in Patients Treated with Immune Checkpoint Inhibitors. Clinical Cancer Research **23**(15), 4270–4279 (2017). https://doi.org/10.1158/1078-0432.CCR-16-3146

[21] Wong, P.F., Wei, W., Smithy, J.W., Acs, B., Toki, M.I., Blenman, K.R.M., Zelterman, D., Kluger, H.M., Rimm, D.L.: Multiplex Quantitative Analysis of Tumor-Infiltrating Lymphocytes and Immunotherapy Outcome in Metastatic Melanoma. Clinical Cancer Research **25**(8), 2442–2449 (2019). https://doi.org/10.1158/1078-0432.CCR-18-2652

[22] Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., Nolan, G.P.: Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences **111**(26), 2770–2777 (2014). https://doi.org/10.1073/pnas.1408792111

[23] Skinnider, M.A., Squair, J.W., Kathe, C., Anderson, M.A., Gautier, M., Matson, K.J.E., Milano, M., Hutson, T.H., Barraud, Q., Phillips, A.A., Foster, L.J., La Manno, G., Levine, A.J., Courtine, G.: Cell type prioritization in single-cell data. Nature Biotechnology **39**(1), 30–34 (2021). https://doi.org/10.1038/s41587-020-0605-1

[24] McInnes, L., Healy, J., Saul, N., Grossberger, L.: Umap: Uniform manifold approximation and projection. The Journal of Open Source Software **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861

[25] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using umap. Nature Biotechnology **37**(1), 38–44 (2019). https://doi.org/10.1038/nbt.4314

[26] Moses, L., Pachter, L.: Museum of spatial transcriptomics. Nature Methods **19**(5), 534–546 (2022). https://doi.org/10.1038/s41592-022-01409-2

[27] Fang, S., Chen, B., Zhang, Y., Sun, H., Liu, L., Liu, S., Li, Y., Xu, X.: Computational approaches and challenges in spatial transcriptomics. Genomics, Proteomics & Bioinformatics **21**(1), 24–47 (2023). https://doi.org/10.1016/j.gpb.2022.10.001

[28] Chen, H., Li, D., Bar-Joseph, Z.: SCS: cell segmentation for high-resolution spatial transcriptomics. Nature Methods **20**(8), 1237–1243

(2023). https://doi.org/10.1038/s41592-023-01939-3

[29] Danaher, P., Kim, Y., Nelson, B., Griswold, M., Yang, Z., Piazza, E., Beechem, J.M.: Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. Nature Communications **13**(1), 385 (2022). https://doi.org/10.1038/s41467-022-28020-5

[30] Liu, Z., Wu, D., Zhai, W., Ma, L.: SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. Nature Communications **14**(1), 4727 (2023). https://doi.org/10.1038/s41467-023-40458-9

[31] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv (2016). https://doi.org/10.48550/arXiv.1607.06450. http://arxiv.org/abs/1607.06450

[32] Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019)