mosna reveals different types of cellular interactions predictive of response to immunotherapies and survival in cancer.

Alexis Coullomb[1, 2], Paul Monsarrat[2, 3, 4], and Vera Pancaldi[1]

[1]CRCT, Université de Toulouse, Inserm, CNRS, Université Toulouse III-Paul Sabatier, Centre de Recherches en Cancérologie de Toulouse, Toulouse, France

[2]RESTORE Research Center, Université de Toulouse, INSERM 1301, CNRS 5070, EFS, ENVT, Toulouse, France

[3]Oral Medicine Department and Hospital of Toulouse – Toulouse Institute of Oral Medicine and Science, CHU de Toulouse, Toulouse, France

[4]Artificial and Natural Intelligence Toulouse Institute ANITI, Toulouse, France

September 23, 2024

Abstract

Spatially resolved omics enable the discovery of tissue organization of biological or clinical importance. Despite the existence of several methods, performing a rational analysis including multiple algorithms while integrating different conditions such as clinical data is still not trivial. To make such investigations more accessible, we developed *mosna*, a Python package to analyze spatial omics data with respect to clinical or biological data and to gain insight on cell interaction patterns or tissue architecture of biological relevance. *mosna* is compatible with all spatial omics methods, it leverages *tysserand* to build accurate spatial networks, and is compatible with Squidpy. It proposes an analysis pipeline, in which increasingly complex features computed at each step can be explored in integration with clinical data, either with easy-to-use descriptive statistics and data visualization, or by seamlessly training machine learning models and identifying variables with the most predictive power.

mosna can take as input any dataset produced by spatial omics methods, including sub-cellular resolved transcriptomics (MERFISH, seqFISH) and proteomics (CODEX, MIBI-TOF, low-plex immuno-fluorescence), as well as spot-based spatial transcriptomics (10x Visium). Integration with experimental metadata or clinical data is adapted to binary conditions, such as biological treatments or response status of patients, and to survival data. We demonstrate the proposed analysis pipeline on two spatially resolved proteomic datasets containing either binary response to immunotherapy or survival data. mosna identifies features describing cellular composition and spatial distribution that

can provide biological insight regarding factors that affect response to immunotherapies or survival.

Availability and implementation: *mosna* is made publicly available to the community, together with relevant documentation and tutorials implemented as Jupyter notebooks to reproduce the result at https://github.com/AlexCoul/mosna

Corresponding authors: alexis.coullomb@inserm.fr, vera.pancaldi@inserm.fr

1 Introduction

The spatial organization of cell types and their interactions play a major role in the development and function of organs and organisms, and in the progression of diseases such as solid tumors. Recent developments in single cell RNA sequencing technologies have enabled the discovery of a rich variety of cell types and states in several animal models or diseases, as well as the uncovering of new signaling pathways of clinical importance. However, the tissue dissociation step required for single-cell sequencing eliminates the information regarding spatial context within solid tissues, thus impairing the discovery of interactions between cell types or states that are potentially relevant to address biological or clinical questions [1]. More recent developments in spatially-resolved omics methods allow quantifying with near single-cell or sub-cellular resolution the abundance of several tens of proteins [2, 3, 4] and up to several thousands of transcripts [5, 6, 7], while preserving the spatial integrity of samples. Given the promises this new type of technologies offer, several methods and packages have been developed to jointly consider the spatial and 'omic' information [8] and to compute statistics or find patterns [9, 10] in the spatial organization of cells. Nevertheless, analyzing the complex datasets produced by these methods in practice can be a daunting task, and extracting biologically or clinically relevant information remains a challenge.

To overcome these difficulties and offer the community tools for integrating any type of spatially-resolved omic experiment into an optimized workflow, we here propose the Multi-Omics Spatial Networks Analysis library (*mosna*), which extracts relevant cellular interactions or tissue spatial organization features and relates them to experimental or clinical conditions. This library includes the computation of assortativity and mixing matrices [11, 12], which quantify preferential interactions between cell subsets in terms of statistical estimates across the whole sample. *mosna* also implements niche discovering methods, including the Neighbors Aggregation Statistics (NAS) [13] and SCAN-IT [14] methods, to discover relevant local cellular neighborhoods by aggregating several features, such as cell types or markers data. These methods define niches across samples, which is essential to compare patients or biological conditions. Furthermore, *mosna* implements methods to foster the analyses and test the relevance of features to the experimental or clinical conditions examined, in particular identifying biomarkers of response to immunotherapy. To demonstrate *mosna*'s ability to discover clinically relevant spatial features, we exemplify the analysis framework applying it to two spatially resolved proteomic datasets, one with binary responder / non-responder clinical data, the second containing continuous right-censored survival data. *mosna* is the first tool proposing an integrated user-friendly and flexible framework to analyze any spatial-omic data with respect to clinical data or biological conditions, starting with the simplest engineered features (cell types proportions) and guiding users towards more complex analyses, including state-of-the

art niche discovering methods. The important step of training machine learning models to gain insight on clinically relevant factors and formulate new hypotheses is also integrated.

2 Methods

2.1 Spatial network reconstruction

To discover potentially clinically relevant cellular interaction patterns, *mosna* leverages the tysserand [15] library to represent tissues as spatial networks. In these networks, nodes are cells for (sub)cellular resolved experiments, or spots for spot-based experiments, and edges indicate physical proximity between nodes. This representation is compatible with any type of spatially resolved omic method, including image-based proteomics (MIBI-TOF[3], t-CyCIF[2], CODEX[4]), classic multi-marker IHC and IF, image-based transcriptomics (MERFISH[5], SeqFISH[6], osmFISH[16]) or sequencing-based spatial transcriptomics (VISIUM 10x[7], Slide-seq[17]). Using tysserand, both 2D and 3D networks can be reconstructed with high speed and accuracy, based on segmented images or cells' coordinates, with several reconstruction methods, including an optimized Delaunay triangulation method for tissue network reconstruction. This network representation is agnostic of the type of measurement, and nodes in these networks can hold arbitrary attributes, such as cell position, raw and processed transcriptomics or proteomics data, cellular or nuclear morphology, or any user-defined feature.

With spots arranged on a regular lattice such as with 10x Visium experiments, nodes are linked with all their neighbors within the minimal inter-spots distance. When reconstructing networks from irregularly arranged nodes, such as with MERFISH experiments, nodes are linked with a dedicated Delaunay-based method for spatial omics data. Briefly, Delaunay triangulation produces virtual cells around nodes, and edges link contacting virtual cells. To clean reconstruction artifacts, especially cells contacting each other while nodes are far apart, while accounting for areas of varying nodes density, *tysserand* performs an adaptative edge trimming. During this step, for each node, edges are ranked by their lengths, and only edges with lengths above twice the 3rd shortest edges are trimmed. At the end of the spatial network reconstruction process, attributes such as user-defined cell-type, histological region or molecular counts, or any other feature are added to nodes, and features computed during the pipeline are progressively added to nodes' attributes. The resulting network data structure can then be used to apply any spatial omics analysis method.

2.1.1 Training of machine learning models on extracted sample features

For binary tasks such as response to therapy, *mosna* makes it straightforward to train a machine learning model, which is an elastic-net penalized logistic regression (LR) model trained with cross-validation, evaluate its performance, and assess the importance of variables in this model by plotting the coefficients weights. For survival data, *mosna* leverages the scikit-survival library to train an Elastic Net-penalized Cox's proportional hazard (CoxPH) model with hyperparameter tuning with cross-validation. The model is trained on variables defined at the sample level from nodes' attributes, such as proportions of cell types or niches, and *mosna* uses the lifelines library to assess the significance of the CoxPH model's coefficients. For each statistically significant variable *mosna* automatically finds the best threshold which maximizes the log-rank test, that is to say *mosna* finds the threshold to define two patient groups with the

highest difference in survival time.

2.1.2 Network assortativity and mixing matrix

To decipher whether cell types interactions or patterns at the scale of whole samples can explain response to therapy, *mosna* can compute the mixing matrix (MM) and the assortativity coefficient (AC) of their corresponding spatial networks. Assortativity is a single number which is a general measure of preferential interactions between nodes that share the same attributes, like cell types or marker positiveness [11, 12], which has also been applied to describe interaction patterns in 3D genome structure [18, 19]. The AC is computed from the mixing matrix (MM), where each element e_{ij} is the fraction of edges between nodes with attribute i and nodes with attribute j (Figure 1b and 3b). After defining a_i and b_j as the marginal sums over row i and column j respectively, this matrix satisfies:

$$\sum_{ij} e_{ij} = 1, \sum_{j} e_{ij} = a_i \,, \sum_{i} e_{ij} = b_j$$

and the AC r is computed from the mixing matrix following:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

On the other hand, MM can also inform us more precisely on how a specific type of cell tends to preferentially interact with another type of cell, considering its off-diagonal elements. The AC and MM of networks are traditionally computed on exclusive attributes, for instance cell types. To our knowledge, *mosna* is the only library able to compute AC and MM on non exclusive attributes, like positiveness to markers. This is particularly interesting to study preferential interactions between cells positive for a pair of markers, when they are also positive for other markers, as is often the case in spatial omics data.

In order to account for differences in mixing patterns that are expected depending on cell type proportions, we compute the z-scored MM and the z-scored AC after random permutations of cell attributes in the network, thus preserving cell type proportions. We have used undirected networks, but depending on the biological question, directed networks can be used to represent asymmetric interactions, *mosna* also integrates the computation of the mixing matrix for directed networks.

2.1.3 Batch correction of omic data

Since technical batch effects can influence downstream analyses from molecular data, *mosna* can perform batch correction directly on molecular data using the scanorama library [20], which was both one of the top performing methods in a recent benchmark [21] and one of the few methods that can directly correct molecular data. When no information is available about batch identifiers, they are defined using patient's identifiers.

2.1.4 Niche definition

In order to discover local cellular communities, also called neighborhoods or niches [22, 23], characterised by specific cell type proportions or marker levels, *mosna* also implements the Neighbors Aggregation Statistics (NAS) method,

which we originally developed to study spatial patterns in cell composition in the mouse cortex[13]. In each sample, on each node and for each specific attribute, we first aggregate values of the node with the values of its first neighbors (Figure 1d). We then compute several statistics for each of these attributes, such as the mean or median, to identify a central tendency, and the standard deviation or interdecile range, to quantify the variations of each variable (marker, cell type, or other) within this neighborhood. By repeating this process for all nodes over all samples in the cohort, we obtain a NAS table with N^C rows and $N^V \times N^S$ columns, where N^C is the number of cells, N^V the number of variables, and N^S the number of statistics computed on aggregated variables. From this table, which contains feature vectors for each cell in each sample, we perform dimensionality reduction and clustering. The resulting clusters are niches, defined either by proportions of cell types or by statistics on marker levels, depending on the analysis choices. The niches can be visualized on the dimensionality reduced projection of the NAS data feature space, and on the samples' spatial networks (Figure 2e and 3e). The process of aggregating variables across neighbors, computing statistics and performing dimensionality reduction (DR) and clustering has been proven fruitful to find microanatomical tissue structures in healthy and diseased lung and cellular niches in tumors, as evidenced by the recent development of the UTAG [10] and CellCharter [9] libraries. As the NAS method proposes an arbitrary number of neighbors order and several options for the statistics, and different methods for the DR and clustering steps, it can be viewed as a generalization of the UTAG and CellCharter methods.

In order to optimize the numerous hyperparameters used in dimensionality reduction and clustering, *mosna* proposes to maximise the performance of prediction of survival or classification. For each parameter set, niches are defined by the NAS method, and for each patient the proportions of cells in niches are used as variables to predict response to therapy. This proportion table can be transformed after normalization by total number of cells, by cell types proportions or by niches, or using a centered log ratio (CLR) transformation. After each training of the machine learning models, performance metrics and data visualization of predictive features are saved for later selection. The performance metrics used for binary classification are the ROC AUC score, suitable for balanced datasets like this one, and bi-clustering plots of patients and niches proportions, while the concordance index and Kaplan-Meier curves are used for survival analysis. The saved figures allow the user to select models with both good prediction performance and sensible bi-clustering or survival predictions. In order to speed-up computation, GPU implementations of these steps are used if a GPU and associated libraries are available, while intermediate results are saved on disk for faster reuse. At the end of the hyperparameter search, users can select the best models to investigate niches predictive of clinical data and their composition. As *mosna* intends to facilitate the use of state of the art spatial omics analysis methods, it can evolve to incorporate new niche discovering methods as they appear. For the moment, it includes the SCAN-IT library [14], which is one of the top performer niche discovering methods according to a recent benchmark [24].

2.2 A rational analysis pipeline for spatial omics and clinical data

Given the great number of types of analyzes and parameters, *mosna* proposes an analysis pipeline in which variables of increasing sophistication are computed in successive steps. The first step is to compute cell types proportions, as even if this variable is not related to spatial information *per se*, it is a very standard measure that has proven predictive of response to therapy in particular diseases [25]. *mosna* can automatically compute all ratios of a set of variables,

defined as $a/(a+b)$ for variables a and b, so the importance of ratios of cell types proportions can be assessed with respect to clinical data. The next computed features are cell types interactions, defined by the mixing matrix of spatial networks. Finally, *mosna* niches across samples defined with the NAS or other implemented methods are used as the model features. At each step, computed variables can be used for descriptive statistics, data visualization (bi-clustering, Kaplan-Meier curves), bi-variate analyses with p-values corrected for false discovery rate with the Benjamini/Hochberg method by default, and training of machine learning models to predict outcome or survival, and to distinguish features of clinical importance. Inspection of models' weights can then inform on the biological processes and help formulating new hypotheses and suggest new analyses.

3 Results

3.1 Patterns in the spatial organization of cells are predictive of response to immunotherapy

We have developed the *mosna* framework to extract quantitative features from spatial omics data (see Methods). To illustrate the step-by-step use of our *mosna* analysis for spatially resolved data, we show an application to a spatially resolved proteomic dataset of Cutaneous T-Cell Lymphoma (CTCL), generated with the CODEX technology on 70 samples from 14 patients treated with anti-PD-1 immunotherapy, where 7 patients responded, and 7 are non-responders [4]. In this CODEX CTCL cohort, none of the cell type proportions had a statistically significant difference of distribution between responders and non-responders (Figure 2a), as noticed by the authors of the original paper describing this dataset, which results in a poor bi-clustering of patients and cell type proportions (Figure SI 1a). After computing the cell types proportions with *mosna*, an elastic-net penalized logistic regression model with cross validation was trained to predict response to therapy. The performance was poor, as the ROC AUC was 0.5 (Figure SI 1b), which is equivalent to a random classifier. This confirms that cell type proportions do not carry information to predict response to immunotherapy in this dataset. To explore the hypothesis that an imbalance in the abundances of specific cell types could explain clinical outcomes, we used *mosna*'s ability for feature engineering to generate various ratios of cell type proportions. However, using all these ratios we also failed to predict clinical response (Figure SI 1c). Only three ratios were statistically significantly different between groups: the ratios of lymphatics & macrophages (M1=M2), lymphatics and vasculature, and lymphatics and macrophages ($M2 > M1$) (Figure 2b). When using these 3 variables only, the bi-clustering provided a proper separation of patients per response group, and a logistic regression model achieved an ROC AUC of 1. Further inspection shows that in this balanced dataset with 50% split between responders and non-responders, these features provide a perfect separation of the two classes, explaining the high level for this metric.

To further explore the spatial information contained in the CODEX CTCL dataset, we used *mosna* to compute assortativity of cell types using the z-scored MM on the 70 samples of the cohort, where elements of the mixing matrix correspond to the preferential interactions between two cell types (see methods). Measured z-scored MM elements were then aggregated per patient, taking the mean over each patient's samples. Four pairs of cell types interactions were found significantly different between response groups (Figure 2c), namely preferential interactions of vasculature with

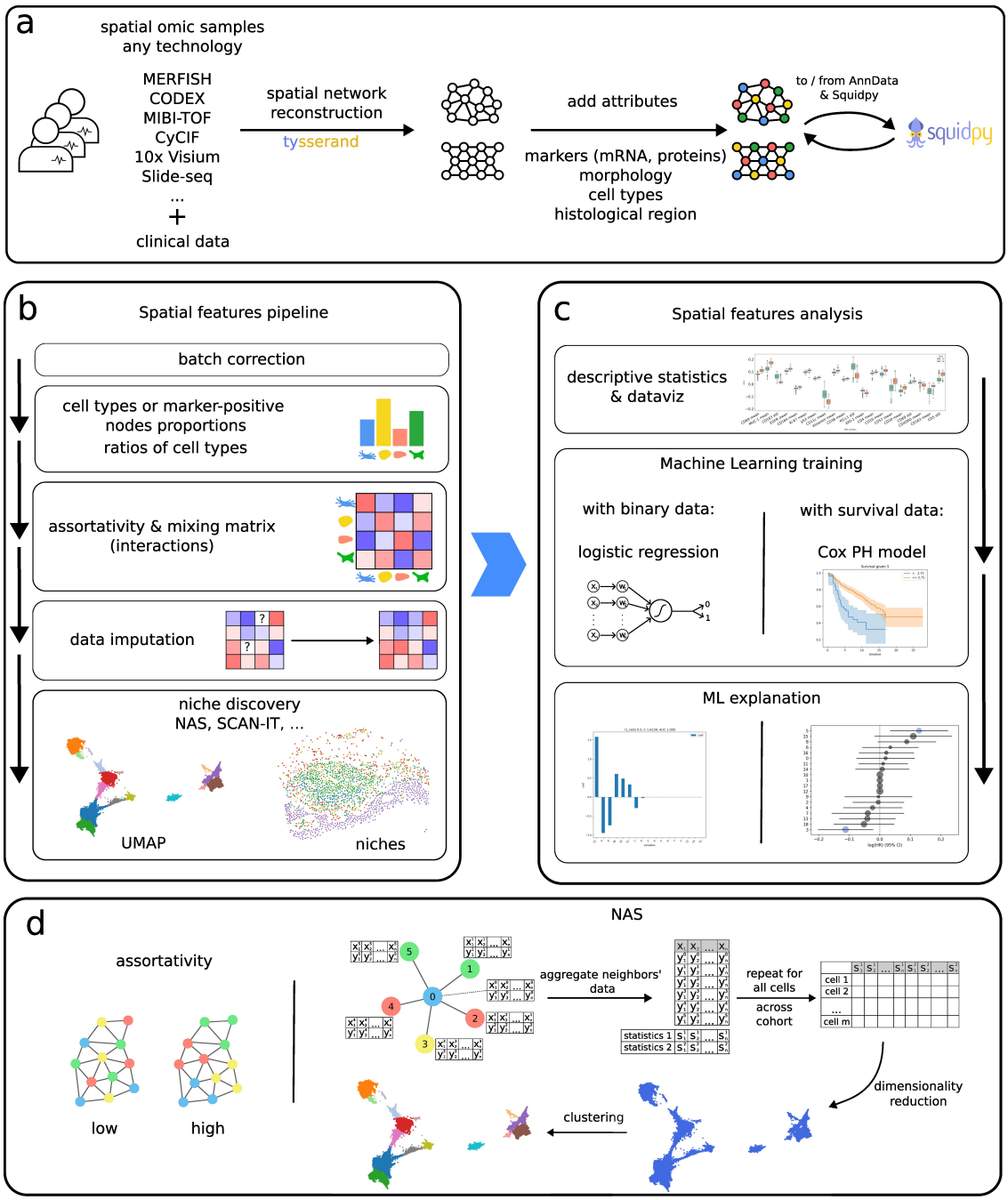

Figure 1: *mosna* analyzes spatial omic data following a rational pipeline. a) Any type of spatial omic data (3D included) can be used with clinical data to reconstruct spatial networks, where nodes hold attributes such as cell types or molecular features. Thanks to its compatibility with the AnnData format and the Squidpy library, the reconstructed networks can be exported to Squidpy objects, or conversely, they can be imported from such objects. b) The analysis pipeline consists of the computation of features with increasing complexity: an initial batch correction can be performed on omic data, then the successive computed features are cell types proportions and their ratios, the assortativity and z-scored interactions between cell attributes, which optionally includes a data imputation method, and, finally, a niche discovering step, performed here with the Neighbors Aggregation Statistics (NAS) method. c) For each type of computed feature, users can easily calculate descriptive statistics, perform data visualisation, and train machine learning models for either a binary classification task or for a survival regression task, using an Elastic Net-penalized logistic regression model or an Elastic Net-penalized survival model respectively. These models can then be inspected by plotting their coefficients, which indicate the most clinically important features. d) *mosna* computes the non-exclusive assortativity on the network, which is a measure of preferential interactions between nodes with similar attributes. A schematic representation of a network with low assortativity (left), and one with high assortativity (right). *mosna* includes the NAS method that defines niches from statistics computed on aggregated omic data around each cell.

itself, macrophages ($M2 > M1$) with lymphatics cells, lymphatics and B cells, and Tregs with CLA$^+$ leukocytes. These features selected as the most differential between the two groups were used to train an LR model that achieved perfect performance. These findings suggest that the global spatial organization of cell types can play an important role in response to immunotherapy. Of note, the computation of the z-scored MM elements is parameter-free, which makes it easy to use and facilitates comparisons across studies.

Phillips et al. defined a *SpatialScore* as the ratio of distances between each $CD4^+$ T cells and their closest tumor cell and their closest Treg. This SpatialScore was then used to stratify patients into responders and non-responders with high accuracy. As this score relates to the interactions between $CD4^+$ T cells and tumor cells of Tregs, we tested whether the ratio of z-scored MM elements between these cell types was linked to response to therapy. We thus leveraged *mosna*'s ability to compute composed variables to compute all ratios of MM elements, and we filtered the ratios that included both the interactions between Tregs and CD4+ T cells and the interactions between tumor cells with any other cell type. Only the ratio 'Tregs - CD4+ T cells / tumor cells, intraepithelial - tumor cells' and 'Tregs - CD4+ T cells / tumor cells - nerves' are significantly different between the response groups, whereas the ratio 'Tregs - CD4+ T cells / tumor cells - CD4+ T cells' corresponding to the Spatial Score is not significantly different between response groups (Fig SI 2a). Without filtering all the computed MM ratio, we noticed that 333 ratios were found significantly different between responders and non-responders ($p - value < 0.05$), the first one being the MM ratio 'macrophages ($M2 > M1$) - B cells Z / tumor cells - macrophages ($M1 > M2$)' (Fig SI 2b).

As evidenced by previous work [26, 27], the mere counts of cell types in a specific tumour sample can be insufficient to predict or explain disease progression or response to therapy, and integration of spatial data may be required. Thus, to explore whether local cell interactions and spatial structures could influence response to therapy, we used *mosna* to look for spatially defined cellular niches, performing hyperparameter search on the NAS method to identify the most predictive niches (see Methods). One top performing set of parameters lead to the definition of 10 niches from the spatial configuration of protein markers data. With *mosna*, one can easily look at the distribution of cell types across niches, normalized either by absolute proportions across samples, by cell type or by niche (Figure 2d left). Some niches are composed mostly of a single cell type (referred to as "pure niches"), whereas other niches present several cell types interacting locally. This niche definition lead to a bi-clustering of patient and niches proportions in the correct patient groups (Figure 2d center), and from these proportions *mosna* could train an LR model predicting response to immunotherapy with high performance (ROC AUC = 1.0) (Figure 2d right). To better understand such a high prediction performance, we checked the distributions of proportion of cells in niches per response group, and found that several niches have a striking difference in distribution between groups (see Fig SI 3a). A PCA computed on the first 3 most predictive niches also showed that groups could be easily separated by a straight line (see Fig SI 3b).

As we inspected various sets of hyperparameters leading to high predictive performance, we could notice that across all models the same 3 niches have high importance in predicting response to immunotherapy. The pure niches of tumor cells and of stroma cells are positively associated with response to therapy, whereas a mixed niche of mostly epithelium, vasculature, Tregs and macrophages (M1=M2), termed "EVTM", is associated with non-response to immunotherapy. Other niches appear frequently as predictive of response or non-response across a number of trained models, such as a mixed niche of mostly macrophages ($M1 > M2$), epithelium, Tregs and tumor cells, termed "METT", associated with

non-response, and a niche made of stroma, tumor cells, macrophages ($M1 > M2$), CD8 cells, CD4 cells, plasma cells, neutrophils and nerves, which is associated with response to therapy.

Of note, at the beginning of the analysis pipeline, proportions of cell types, including stroma and tumor cells, were not predictive of response. The fact that niches made mostly of these cell types are predictive of response to immunotherapy suggests that local areas appearing to exclude other cell types are prognostic markers of a favorable response to such treatment. One hypothesis could be that responders are patients with immune cells excluded from tumor regions, and immunotherapy breaks this exclusion, produces the infiltration of immune cells in tumor regions leading to response.

Inspection of niche composition with the selected hyperparameters showed that niches 16, 11 and 13 are similar, as they are mainly composed of stroma cells, though only niches 16 and 11 are informative of response to therapy. We thus investigated why 3 niches that are seemingly similar in cell types composition were defined, and only 2 of them were predictive of response to therapy. A more detailed inspection showed that niches 11, 13 and 16 were made of 6%, 0.5% and 0% of tumor cells respectively. Whereas the content of tumour cells could explain a difference between niche 11 and the remaining two, the difference in cell types composition between niches 13 and 16 seemed too faint to explain a segregation of these local neighborhoods and their predictive power. We thus performed differential niche analysis, in which neighbors aggregated statistics obtained with the NAS method are compared between niches. From the niches 13 and 16 we compared only stroma cells and their NAS variables (Figure 2f), hypothesising that the amount of tumor cells in niche 13 is so low that differences should be due to a different environment (chemokines, mechanics, ...) and not "leakage" of tumor cell's data into neighborhoods. Between these 2 niches, 102 NAS variables were found having statistically different distributions after FDR correction, with stroma cells of niche 16 having higher levels of CD69 and MUC-1, and lower levels of CD163 than stroma cells of niche 13. Interestingly, performing the differential niche analysis on the NAS variables shows that what changes between niches is the variability in marker levels, with for example cell neighborhoods in niche 16 presenting a more homogeneous level of expression of PD-L1 than in niche 13, while levels of CD69 and CD5 have higher variablility.

3.2 Survival analysis in breast cancer and associated factors

To exemplify *mosna*'s ability to discover clinically relevant spatial features with respect to survival data, we applied it to a dataset generated by Danenberg et al. [28] with Imaging Mass Cytometry on 693 breast tumors samples. The authors linked spatial omics data with genomic and clinical data, and discovered conserved tumor microenvironment (TME) structures across various cancer subtypes. Here we tested whether starting from cellular molecular markers we could find some spatial structures that would be predictive of survival, even without clinical or genomic *a priori*.

As a first step of the pipeline, a CoxPH model was trained on cell type proportions. Although none of the cell types proportions are significantly associated with longer survival, cells defined as 'macrophages ' and 'CD4+ T cells & APCs' tends to be positively associated with longer survival (Figure 3a, right), which is consistent with the literature [29]. Following the analysis pipeline, we then computed the MM, whose elements indicate preferential interactions or avoidance between cell types (Figure 3b). Given the high imbalance in cell type proportions, with some cell types representing less than 1% of cells, several MM elements had undefined values for most of the samples. We thus used

mosna's imputation method to infer the most likely values, where variables were defined for at least 50% of samples, and we deleted variables and samples with too many undefined values.

The resulting MM coefficients were used to train a CoxPH model, and 8 interactions were found to be significantly associated with survival: the interactions between '$FibroblastsFSP1^+$' and '$CK8-18^{hi}CXCL12^{hi}$', and between $CK^+CXCL12^+$ and '$CK8-18^{hi}CXCL12^{hi}$ were associated with longer survival, whereas other interactions were associated with worst prognosis, such as those between myofibroblasts and themselves, between myofibroblasts and macrophages, between $EpKi67^+$ and $CK8-18^{hi}CXCL12^{hi}$, $Myofibroblasts$ and $CK8-18^{hi}CXCL12^{hi}$, $FibroblastsFSP1^+$ and $CK^+CXCL12^+$, and between $CK^+CXCL12^+$ and $CD4^+Tcells$.

As Danenberg et al. defined 10 TME structures found in several breast cancer subtypes, we tested whether *mosna* could find such TME structures, or niches, that are predictive of survival. Given the high number of samples and cells, we chose to leverage the NAS method's flexibility to directly cluster cells using neighbors aggregated statistics, using this time cell types proportions in the neighborhood, skipping the dimensionality reduction step. After the NAS hyperparameters search, the best performing model according to the concordance index had two significant coefficients, corresponding to niches 3 and 5, that were respectively positively and negatively associated with survival (Figure 3c right). Given the high cell type proportion imbalance, the matrix plot of cell types proportions per niche proved to be poorly informative (Figure 3c center). We thus performed differential niche analysis of NAS variables, between each of the significant niches versus all other niches (Figure 3d). Niche 3 consists of neighbors made up of more myofibroblasts, fibroblasts and endothelial cells, and less $CK^{med}ER^{lo}$ and $CK^{lo}ER^{lo}$ than other niches, which corresponds to the 'vascular stroma' niche found by Danenberg et al. and associated with longer survival. In contrast, niche 5 consists of neighborhoods with more $HER2^+$ cells, whereas the composition of niche 5 does not seem to be enriched in $HER2^+$ cells with respect to other niches. This is verified by a detailed inspection of niche 5 composition, which includes 2.2% of $HER2^+$ cells, knowing that this cell phenotype represents 3.2% of all cells in the studied cohort, whereas we observe a much higher proportion of those cells in the neighborhood of niche 5 cells, up to 100%, as shown in the histogram Figure SI 4. To understand these seemingly contradictory results, we should here highlight the distinction between the proportion of cell types in a specific niche, and the proportion of cell types in the *neighborhood of cells* assigned to a niche. Visual inspection performed jointly on the spatial organization of cell phenotypes and niches (Figure 3e) shows that non-$HER2^+$ cells neighboring $HER2^+$ cells are often assigned to niche 5, with $HER2^+$ cells forming continuous areas with preferential interactions between themselves. The presence of $HER2^+$ cells being linked to shorter survival is consistent with the $HER2^+$ mutation being linked to breast cancer aggressiveness [30], though the organisation of $HER2^+$ cells in continuous neighborhoods has not been observed before, to the best of our knowledge.

4 Discussion

Spatially resolved omics data represent an opportunity to understand how cell interactions and spatial factors influence tissue functions in health or disease progression, but the analysis of such complex datasets with respect to clinical data can be challenging. Here we present the Multi-Omics Spatial Networks Analysis library (*mosna*), which facilitates the analysis of any type of spatial omics data in relation to clinical data or biological conditions. *mosna* proposes an

analysis pipeline in which features of increasing complexity are computed and used for data visualisation, calculation of descriptive statistics and training of machine learning models. The progression in complexity allows biologists to start from the simplest hypotheses (disease progression is related to cell types proportions) to more complex ones (disease progression is related to cellular interactions or their spatial organization).

The computation of cell type interactions has been introduced in previous studies [31, 8], but *mosna* additionally allows to compute interactions between cells with non-exclusive attributes, such as positiveness for different markers. In order to account for imbalance in attributes proportions, *mosna* performs multiple randomizations of assignation of attributes to cells to compute z-scored interaction scores. This z-score computation relies on the hypothesis that the distribution of randomized interactions follows a normal distribution, which has never been tested. We believe that in the future the development of an empirical score accounting for the appropriate distribution, or even the development of a theoretical distribution of interactions given attributes proportions, would lead to a more accurate estimation of scores of preferential interactions between cell types.

Finally, *mosna* can find local neighborhoods, or niches, using user-defined cell types or any other attribute, such as markers data or cell shape descriptors. To mitigate the risk of identifying niches that are specific to patients and hence cannot generalise as predictive features, *mosna* integrates the scanorama batch correction method, which can directly correct raw markers data. *mosna* integrates the Neighbors Aggregation Statistics (NAS) method, initially developed to discover niches in seqFISH data, and demonstrated here on two spatial proteomic datasets generated by CODEX and IMC. As the NAS method allows to aggregate data up to any order of neighbors, vary dimensionality reduction parameters, and change the clustering method, it can be viewed as a generalization of the CellCharter and UTAG niche discovering methods. This generalization comes at a cost though: users have to choose parameters in a high dimensional space. To select the best hyperparameters, *mosna* includes a hyperparameter optimization step, in which predictive models are trained with cross-validation and Elastic Net regularization, and the statistical significance of coefficients are computed for CoxPH models to limit the risk of over fitting. This hyperparameter search is facilitated by leveraging GPU implementations of dimensionality reduction and clustering methods and saving intermediate results on disk in a rational structure for successive reuse. Furthermore, the discovery of niches is performed jointly on all samples, which allows the identification of rarer or smaller neighborhoods and spatial patterns that occur across images, facilitating the interpretation of niches. The definition of niches across samples is a requirement to train predictive models, whether it is to estimate the survival of patients or their response to therapy. *mosna* already integrates the SCAN-IT method and will keep integrating other niche discovering methods to ease their use in the Python ecosystem and facilitate analysis with respect to clinical data. Currently *mosna* can create and modify AnnData / Squidpy objects [8] and retrieve data from them, but further integration with these libraries could be developed in the future. Another planned extension of *mosna* is the integration of other machine learning models. Currently *mosna* implements Elastic Net-penalized logistic regression and CoxPH models, but other types of models such as gradient boosting trees or neural network-based models, both for classification and survival regression, could be beneficial to users.

We demonstrated how *mosna* could extract relevant information from spatial omics datasets generated by different technologies. These analyses showed that in those datasets, cell types proportions were not predictive variables, whereas mosna can find clinically relevant cell types interactions and niches. To provide a biological interpretation of the results,

the user can explore the differences in cell type composition, NAS variables or raw markers between these niches. The interpretation of these variables is a crucial task to develop more efficient patient stratification algorithms for personalized medicine, and to discover targetable biological pathways and guide the development of novel therapies. We believe that mosna will be a valuable tool for biological and clinical advancements exploiting the wealth of spatially resolved data that is increasingly available in both research and clinical settings.

Key points:

- *mosna* facilitates the analysis of spatially-resolved omic data in integration with clinical data and discovers spatial patterns predictive of response to therapy or survival.
- *mosna* proposes a rational analysis pipeline to test hypotheses of increasing complexity.
- *mosna* including the computation of preferential cellular interactions and the NAS method, a general framework for niche discovery.
- *mosna* facilitates the training of machine learning models to perform integration with clinical data or biological conditions, while features can be inspected to gain insight on biological processes and formulate data-driven hypotheses.

Funding

This work was funded by INSERM; Fondation Toulouse Cancer Santé and Pierre Fabre Research Institute as part of the Chair of Bioinformatics in Oncology of the CRCT and by a JANSSEN Horizon grant to AC and VP. AC acknowledges support from NIH NIMH (1RF1MH128867). This study has been partially supported through the grant EUR CARe N°ANR-18-EURE-0003 in the framework of the Programme des Investissements d'Avenir and the national infrastructure "ECELLFrance: Development of mesenchymal stem cell based therapies" (PIA-ANR-11-INBS-005).

References

[1] Yingcheng Wu et al. "Spatial omics: Navigating to the golden era of cancer research". In: Clinical and Translational Medic 12.1 (2022), e696.

[2] Jia-Ren Lin et al. "Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes". In: elife 7 (2018).

[3] Leeat Keren et al. "A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging". In: Cell 174.6 (2018), pp. 1373–1387.

[4] Darci Phillips et al. "Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma". In: Nature communications 12.1 (2021), p. 6726.

[5] Chenglong Xia et al. "Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression". In: Proceedings of the National Academy of Sciences 116.39 (2019), pp. 19490–19499.

[6] Chee-Huat Linus Eng et al. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: Nature 568.7751 (2019), pp. 235–239.

[7] Patrik L Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". In: Science 353.6294 (2016), pp. 78–82.

[8] Giovanni Palla et al. "Squidpy: a scalable framework for spatial omics analysis". In: Nature methods 19.2 (2022), pp. 171–178.

[9] Marco Varrone et al. "CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity". In: Nature Genetics 56.1 (2024), pp. 74–84.

[10] Junbum Kim et al. "Unsupervised discovery of tissue architecture in multiplexed imaging". In: Nature methods 19.12 (2022), pp. 1653–1661.

[11] Mark EJ Newman. "Assortative mixing in networks". In: Physical review letters 89.20 (2002), p. 208701.

[12] Mark EJ Newman. "Mixing patterns in networks". In: Physical review E 67.2 (2003), p. 026126.

[13] Kim-Anh Lê Cao et al. "Community-wide hackathons to identify central themes in single-cell multi-omics". In: Genome biology 22 (2021), pp. 1–21.

[14] Zixuan Cang et al. "SCAN-IT: Domain segmentation of spatial transcriptomics images by graph neural network". In: BMVC: Proceedings of the British Machine Vision Conference. British Machine Vision Conference. Vol. 32. NIH Public Access. 2021.

[15] Alexis Coullomb and Vera Pancaldi. "Tysserand—fast and accurate reconstruction of spatial networks from bioimages". In: Bioinformatics 37.21 (2021), pp. 3989–3991.

[16] Simone Codeluppi et al. "Spatial organization of the somatosensory cortex revealed by osmFISH". In: Nature methods 15.11 (2018), pp. 932–935.

[17] Robert R Stickels et al. "Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2". In: Nature biotechnology 39.3 (2021), pp. 313–319.

[18] Vera Pancaldi et al. "Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity". In: Genome biology 17 (2016), pp. 1–19.

[19] Miguel Madrid-Mencía et al. "Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks". In: Nucleic Acids Research 48.8 (2020), pp. 4066–4080.

[20] Brian Hie, Bryan Bryson, and Bonnie Berger. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama". In: Nature biotechnology 37.6 (2019), pp. 685–691.

[21] Malte D Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: Nature methods 19.1 (2022), pp. 41–50.

[22] Caleb R Stoltzfus et al. "CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues". In: Cell reports 31.3 (2020).

[23] Yury Goltsev et al. "Deep profiling of mouse splenic architecture with CODEX multiplexed imaging". In: Cell 174.4 (2018), pp. 968–981.

[24] Zhiyuan Yuan et al. "Benchmarking spatial clustering methods with spatially resolved transcriptomics data". In: Nature Methods (2024), pp. 1–11.

[25] Daniela Bruni, Helen K Angell, and Jérôme Galon. "The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy". In: Nature Reviews Cancer 20.11 (2020), pp. 662–680.

[26] Sebastian Lundgren et al. "Quantitative, qualitative and spatial analysis of lymphocyte infiltration in periampullary and pancreatic adenocarcinoma". In: International Journal of Cancer 146.12 (2020), pp. 3461–3473.

[27] Lili Cao et al. "M2 macrophage infiltration into tumor islets leads to poor prognosis in non-small-cell lung cancer". In: Cancer management and research 11 (2019), p. 6125.

[28] Esther Danenberg et al. "Breast tumor microenvironment structures are associated with genomic features and clinical outcome". In: Nature genetics 54.5 (2022), pp. 660–669.

[29] Suchit Jhunjhunwala, Christian Hammer, and Lélia Delamarre. "Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion". In: Nature Reviews Cancer 21.5 (2021), pp. 298–312.

[30] Sandra M Swain, Mythili Shastry, and Erika Hamilton. "Targeting HER2-positive breast cancer: advances and future directions". In: Nature reviews Drug discovery 22.2 (2023), pp. 101–126.

[31] Denis Schapiro et al. "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data". In: Nature methods 14.9 (2017), pp. 873–876.

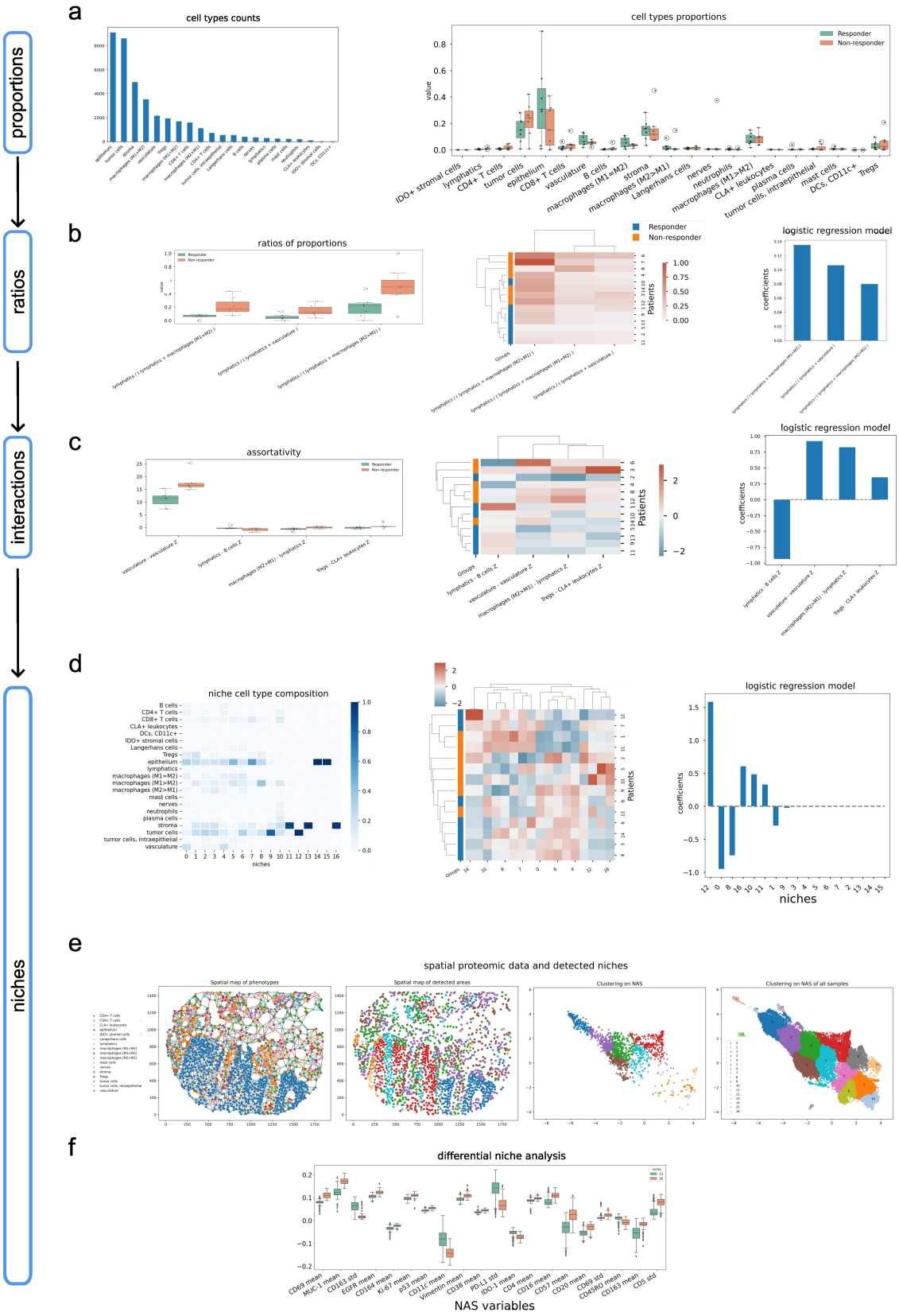

Figure 2: Analysis of spatial omic data with respect to binary response to therapy. a) Cell types proportions (left) and distributions (right) between responders and non-responders. b) Ratio of cell types are used to: visualize their distributions between response groups (left), perform bi-clustering of patients (center) and train a logistic regression model and inspect its coefficients (right). The same analyses are performed for the computed preferential interactions (c) and for niches defined with the NAS method (d), with an additional visualisation of niches composition (left). e) From left to right: spatial organisation of cell types, of niches, and UMAP representation of nodes and their NAS variables, colored by niches, for nodes in the displayed sample, and for the whole dataset. f) Differential niche analysis comparing NAS variables of stroma cells between niches 3 and 16.

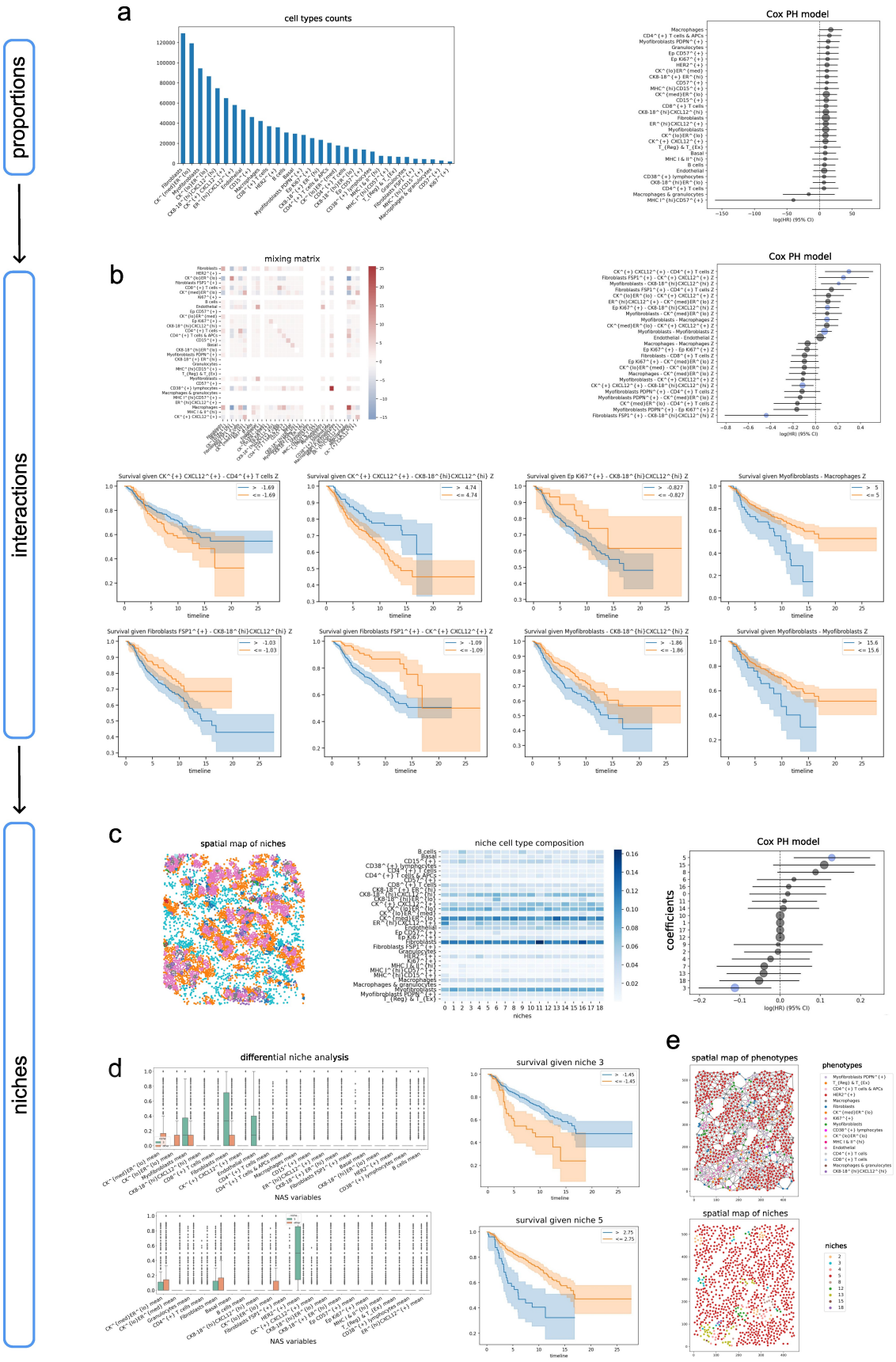

Figure 3: Analysis of spatial omics data with respect to survival data. a) Cell types proportions (left) and coefficients of the CoxPH model trained on these features, with 95% confidence intervals. b) Mixing matrix showing preferential cell type interactions or avoidance for one sample (top left), coefficients of a CoxPH model trained on these features (top right), and Kaplan-Meier curves with patient groups defined from an optimized threshold for each significant variable in the CoxPH model (last two rows) c) Exemplary niche visualization for a sample (left), matrix plot of niche composition (center) impaired by a high imbalance in cell types proportions, and coefficients of the CoxPH model trained on these features (right). d) Differential niche analysis showing the distribution of cell types proportions in neighborhoods for niche 3 versus all other niches (top left) and for niche 5 versus all other niches (bottom left), and Kaplan-Meier curves defined from the corresponding niches (right). e) Spatial organization of cell types (top) and niches (bottom) in a sample.