SpaCeNet: Spatial Cellular Networks from omics data

Niklas Lück[1], Robert Lohmayer[2,3], Stefan Solbrig[2], Stefan Schrod[1], Tina Wipfler[2], Katherine H. Shutta[4,5], Marouen Ben Guebila[4], Andreas Schäfer[2], Tim Beißbarth[1,6], Helena U. Zacharias[7], Peter J. Oefner[8], John Quackenbush[4], and Michael Altenbuchinger*[1]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Germany.
[2]Institute of Theoretical Physics, University of Regensburg, Germany.
[3]Leibniz Institute for Immunotherapy, 93053 Regensburg, Germany.
[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
[5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[6]Campus Institute Data Science (CIDAS), University of Göttingen, Germany.
[7]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany.
[8]Institute of Functional Genomics, University of Regensburg, Regensburg, Germany.

September 1, 2022

Advances in omics technologies have allowed spatially resolved molecular profiling of single cells, providing a window not only into the diversity and distribution of cell types within a tissue, but also into the effects of interactions between cells in shaping the transcriptional landscape. Cells send chemical and mechanical signals which are received by other cells and transduced through signaling cascades from the membrane to the nucleus, where they can subsequently initiate context-specific gene regulatory responses. These interactions and their responses shape the individual molecular phenotype of a cell in a given microenvironment. RNAs or proteins measured in individual cells together with the cells' spatial distribution provide invaluable information about these mechanisms and the regulation of genes beyond processes

*correspondence: michael.altenbuchinger@bioinf.med.uni-goettingen.de

occurring independently in each individual cell. "SpaCeNet" is a method designed to elucidate both the intracellular molecular networks (capturing how molecular variables affect each other within the cell) and the intercellular molecular networks (how cells affect molecular variables in their neighbors). This is achieved by estimating conditional independence relations between captured variables within individual cells and by disentangling these from conditional independence relations between variables of different cells.

1. Introduction

Measurements of spatially resolved RNA or protein expression patterns open unprecedented opportunities to study questions in areas such as developmental biology and pathophysiology, where interactions between cells are known to influence a wide range of processes. Methods such as "spatial transcriptomics" [Ståhl et al., 2016] and Slide-Seq [Rodriques et al., 2019] used molecular barcoding to count mRNAs aggregated in small regions and integrated those with images to produce a transcript map. A drawback of these early methods was that mRNAs from multiple cells in a small region could contribute to the observed signal, masking differences between cells and cell-cell interactions. *In situ* hybridization or sequencing methods can measure the expression of many genes with single-cell resolution (and even subcellular resolution) [Chen et al., 2015, Shah et al., 2016, Lee et al., 2014], but these methods require complex instrumentation and long imaging times. Three-dimensional intact tissue sequencing in single cells has been achieved by STARmap (spatially-resolved transcript amplicon readout mapping) [Wang et al., 2018], which is capable of measuring simultaneously the expression of about 1000 genes at single-cell resolution. More recently, sci-Space has allowed single-cell RNA sequencing, but it is only capable of capturing a small proportion of cells in a tissue [Srivatsan et al., 2021]. Although spatial omics technologies are in their infancy, the launch of commercial products such as the 10X Genomics Visium platform has led to increased interest in methods that will allow the analysis and interpretation of data generated [Svensson et al., 2018, Arnol et al., 2019, Bergenstråhle et al., 2020].

Cells organize themselves spatially within tissues and organisms in order to carry out specific functions. This organization is orchestrated via signals that include both physical interactions via cell-cell contact, chemical signals, and even exosome-mediated transfer of RNAs between cells. Each cell's individual phenotype together with its location in space relative to other cells captures information about this process. For instance, genes encoding chemokines (chemoattractant cytokines that facilitate intercellular communication) are first transcribed and then translated before the respective proteins pass through the cell membrane into the extracellular domain where they recruit leukocytes. When these signals are received by the leukocytes,

they initiate signaling cascades that finally induce a molecular response leading the cells to adapt their individual molecular phenotype and alter their behavior. While such interactions are known to be essential for many biological processes [Armingol et al., 2021], there are no well-established statistical methods to investigate the relationships between spatial organization, gene or protein expression, and cellular phenotype.

The statistical inference of correlation-based molecular networks from high-dimensional omics data is based on the assumption that coordinated expression in a cell might provide insight into processes that are activated or inactivated in different phenotypes. Early attempts used pairwise measures of gene-gene coexpression such as mutual exclusivity, Pearson's and Spearman's correlations and identified network edges between genes based on a correlation threshold [Margolin et al., 2006, Marbach et al., 2012]. Such measures of coexpression can provide insight into active biological mechanisms, but they are vulnerable to spurious associations. These associations can be the consequence of indirect dependencies that cannot be resolved if pairwise relationships between molecular variables are considered only in isolation from all other molecular variables. This stimulated research in high-dimensional statistics, specifically in Probabilistic Graphical Models (PGMs). PGMs resolve the dependency structure of molecular variables and, thus, disentangle direct from indirect associations. This is even possible in high-dimensional settings, where the number of variables is larger than the number of observations. The underlying concept is conditional independence. Two variables X and Y are considered as conditionally independent (given all remaining variables), if X does not provide any additional information about Y that is not already covered by the remaining variables. Thus, although X and Y might be pairwise correlated, they can still be conditionally independent; the pairwise correlation could be just the consequence of X's and Y's indirect relationship mediated via other variables and not due to a direct relationship. This powerful concept made PGMs one of the favored approaches to resolve molecular networks from molecular data [Marbach et al., 2012]. It is nevertheless noteworthy that PGMs are undirected graphical models and do not resolve causal relationships, although, intriguingly, lower bounds on causal effects can be provided [Maathuis et al., 2009].

SpaCeNet is a method for analyzing patterns of correlation in spatial transcriptomics data by extending the concept of conditional independence to spatially distributed information, facilitating reconstruction of both the intracellular and the intercellular interaction networks with single-cell spatial resolution. SpaCeNet was developed to address the diversity of cellular interactions and the various length scales over which intercellular communication occurs (for example, both through direct physical contact and through long-range paracrine signaling). SpaCeNet introduces flexible interaction potentials in combination with appropriate regularization strategies to allow this diversity in both cellular state and spatial communication to be handled ef-

fectively. We validate SpaCeNet in extensive simulation studies and illustrate its capacity to augment exploratory data analysis of spatial transcriptomics data from the mouse visual cortex and the *Drosophila* blastoderm.

2. Results

2.1. Modeling cell-cell interactions

SpaCeNet is built on Gaussian Graphical Models (GGMs), an instance of PGMs which assumes multivariate normal data. GGMs have become a standard for molecular network inference based on observational data (see, for example, the sparse regression-based approaches described by Marbach and colleagues [Marbach et al., 2012] or the comparisons conducted by Ghanbari and colleagues [Ghanbari et al., 2019]). GGMs have been readily adapted to the analysis of single-cell data via non-paranormal transformations [Zhang et al., 2018, Weighill et al., 2021]. Usually, conditional independence (CI) relations are defined for molecular variables X^a and Y^a, measured within one cell a, conditioned on all remaining variables of the same cell a. Assuming that the same set of space-agnostic CI relations applies to all measured cells, this can be expressed as $X \perp Y|\{\text{rest}\}$, in accordance with the traditional nomenclature. Here the set "rest" refers to all variables measured in one cell except for X and Y.

To extend this language to cellular networks, we introduce intracellular and intercellular spatial conditional independence (SCI) relations. Intracellular SCI relations between variables X^a and Y^a measured within one cell a are expressed as $X^a \perp Y^a|\{\text{rest}\}$ with the term "rest" now referring to all other variables of cell a and to all variables of all other cells. We write intercellular SCI relations between variables X^a and Y^b measured in different cells a and b, conditioned on all remaining variables, as $X^a \perp Y^b|\{\text{rest}\}$ with $a \neq b$. Here, the set "rest" refers to all variables of cell a except X^a, all variables of cell b except Y^b, and all variables of all remaining cells. The intra- and intercellular SCI relations captured by SpaCeNet are illustrated in Figure 1. SpaCeNet parameterizes the intercellular SCI relations via spatial interaction potentials, as outlined in the following and as described in detail in the Methods Section 4.1.

We denote potentials that mediate the interaction between the molecular variables X^a and Y^b in cells a and b, respectively, as $\rho_{XY}(r_{ab})$, where r_{ab} is the distance between cells a and b. The parametrization of SpaCeNet is chosen such that if and only if $\rho_{XY}(r_{ab}) = 0$ for all r_{ab}, then $X^a \perp Y^b|\{\text{rest}\}$ for all pairs of cells $a \neq b$. Thus, the SCI relations can be directly obtained from the joint probability density function, see Eq. (1). Importantly, the function $\rho_{XY}(r_{ab})$ depends only on the distance r_{ab} between two cells and thus is identical across all modeled cells. One might ask whether such a one-function-fits-all approach is appropriate to model complex cellular communication

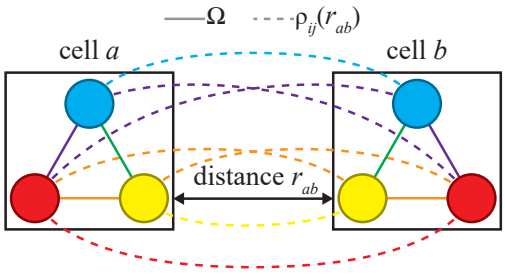


Figure 1: Illustration of intracellular and intercellular Spatial Conditional Independence (SCI) relations captured by SpaCeNet. Circles of the same color represent the same molecular variable in different cells. Solid lines represent edges of intracellular networks, whereas the dashed lines represent the intercellular network edges between cells a and b, which are mediated through the interaction potentials $\rho_{ij}(r_{ab})$. The latter parameterizes the interaction between variables i and j between two cells with distance r_{ab}. The missing dashed green lines imply SCI between the blue and yellow variable (i.e., the corresponding $\rho_{ij}(r_{ab}) = 0$), while SCI does not apply between the yellow and red variable (i.e., $\rho_{ij}(r_{ab}) \neq 0$).

between molecularly diverse cells, considering that even related cell types, such as T-cells and B-cells, fulfill very different tasks and are expected to send and receive very different signals. As briefly illustrated, SpaCeNet can encode such diverse mechanisms. Starting from the full density, Eq. (1), and marginalizing over all variables except X^a and Y^b, the resulting distribution corresponds to the regression formula $X^a = \mu_X - \frac{1}{\omega_{XX}}\rho_{XY}(r_{ab})(Y^b - \mu_Y) + \epsilon$ (see also Eq. (9)). Thus, the effect that molecular variable Y^b in cell b exerts on X^a via cell-cell interaction is directly related to the individual molecular profile of cell b, which means that the molecular phenotype of a cell determines how it communicates with other cells.

A key step in fitting the SpaCeNet model is determining $\rho_{XY}(r_{ab})$, for which we assume no particular functional dependence on r_{ab}. Infinitely separated cells cannot communicate through the release or absorbance of particular signaling molecules and the drop in concentration of these molecules with distance suggests that $\rho_{XY}(r_{ab})$ should vanish for $r_{ab} \to \infty$. This motivates a series expansion of ρ_{XY} in powers of $1/r_{ab}$ (accompanied with an additional factor that smoothes the divergence for $r_{ab} \to 0$, see Methods Section 4.1.2). Parameter estimation was then performed using a pseudo-log-likelihood approach (see Methods Section 4.1.3).

To show that SpaCeNet can reconstruct diverse cell-cell interactions, we performed four illustrative simulations shown in Figure 2a to d, where we simulated radial dependencies corresponding to long-distance interactions (such as via paracrine signaling, Fig. 2a and b), short-distance interactions (such as via cell-cell contact) using an exponentially decreasing potential

with short range (Fig. 2c), and an interaction where the potential first grows with r_{ab}, peaks at average distances, and then goes to zero for large r_{ab} (Fig. 2d). The precise radial dependencies are provided in the caption of Fig. 2. In principle, SpaCeNet can model interaction potentials to arbitrary orders in $1/r_{ab}$. We present the corresponding radial dependencies estimated via SpaCeNet for expansions up to order $(1/r_{ab})^L$ for $L = 1$ (solid lines), $L = 3$ (dashed lines), and $L = 10$ (dotted lines). As expected, the approximation of the true underlying functional dependencies improves with increasing L. However, increasing the order in $1/r_{ab}$ results in higher model complexity, which in turn comes at the cost of additional parameters, increasing the risk of overfitting and computational burden.

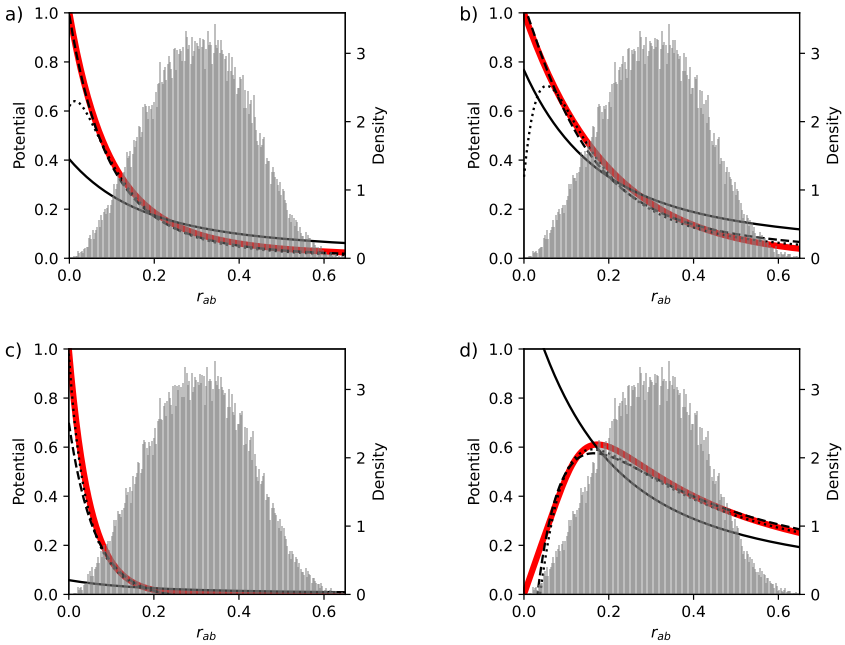


Figure 2: Study to assess the effect of the expansion order L. Different potentials (red lines) have been used to simulate data, and models with $L \in \{1, 3, 10\}$ have been fitted to these data (represented by the black solid, dashed, and dotted lines, respectively): (a) a standard decreasing potential corresponding to the second-order term in the series expansion, $\rho(r_{ab}) = (1 - \exp(-r_{ab}/r_0))^2 (r_0/r_{ab})^2$ with $r_0 = 1/10$, (b) a long-range exponential potential $\rho(r_{ab}) = \exp(-5\, r_{ab})$, (c) a short-range exponential potential $\rho(r_{ab}) = \exp(-20\, r_{ab})$, and (d) a potential $\rho(r_{ab}) = 10\,(1 - \exp(-r_{ab}/2))\,(1 - \exp(-1/(5r_{ab})^2))$ which increases for small r_{ab} and then decreases again. The left y-axis corresponds to the values of the potentials and the right y-axis to the density of the pairwise distances r_{ab} shown as histogram in the background of the figures. For details about data simulation see Methods Section 4.2.

2.2. Simulation studies

SpaCeNet solves a high-dimensional optimization problem in which the number of parameters can be of the same order of magnitude as the number of cells, or even larger. This makes model regularization mandatory [Schäfer and Strimmer, 2005a, Altenbuchinger et al., 2020]. To control overfitting and to achieve parsimonious intra- and intercellular networks, we added regularization terms to the objective function Eq. (10) which induce sparseness in the intracellular network edges (via lasso regularization) and sparseness in the intercellular network edges (via group-lasso regularization). For details see Methods Section 4.1.4.

We performed four simulation studies to analyze the performance of SpaCeNet in recovering intra- and intercellular network edges. To this end, we generated artificial data from the full probability density Eq. (1), where we simulated cell-cell interaction potentials $\rho_{ij}(r_{ab}) = \Delta\rho_{ij} \exp(-\phi_{ij} r_{ab})$ that exponentially decrease with cell-cell distance r_{ab}. We simulated n cells per $S \in \{1, 10, 100\}$ replicates and present our results as a function of the total number of cells $n \cdot S \in \{10^3, 10^4, 10^5\}$ in Fig. 3.

In a first simulation study, we used a comparatively large range of $1/\phi_{ij} = 1/5$ (see Methods Section 4.2). Top and bottom plots of Fig. 3a give the performance in terms of the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC), respectively, to recover the correct intracellular networks (the intracellular precision matrix $\mathbf{\Omega}$, left figure) and the intercellular interaction networks (the cell-cell interaction parameters $\mathbf{\Delta\rho}$, right figure). We find that SpaCeNet is capable of reconstructing intracellular networks across the full range of simulated cells, $n \cdot S$, with an AUROC and AUPRC larger than 0.76 (left Figs. 3a). Edge recovery of cell-cell interactions was achieved with a median AUROC ranging from 0.530 for $n = 10^3$ and $S = 1$ (blue) to 0.997 for $n = 10^3$ and $S = 100$ (green). It is worth noting that edge recovery for both the intra- and intercellular networks was slightly better if cells are distributed across measurements, as seen by comparison of the AUROC for $n = 10^3$ and $S = 1$ (blue), for $n = 100$ and $S = 10$ (orange), and for $n = 10$ and $S = 100$ (green), that all share the same total number of cells $n \cdot S = 10^3$. This trend also persisted for $n \cdot S = 10^4$ and AUPRC (Fig. 3a, bottom).

We also tested the same class of potentials with medium range $1/\phi_{ij} = 1/10$ (Fig. 3b) and short range $1/\phi_{ij} = 1/20$ (Fig. 3c). The results were similar to the first simulation study. The edge recovery in terms of AUROC and AUPRC improved with an increasing number of cells $n \cdot S$, the more so when cells were distributed across multiple samples S.

Finally, we studied how diverse cell-cell interactions impair model inference (Fig. 3d). For this, we repeated the previous simulations, but created flexible potentials $\rho_{ij}(r_{ab}) = \Delta\rho_{ij} \exp(-\phi_{ij} r_{ab})$ with $\phi_{ij} \sim \text{Unif}(5, 20)$ and $\Delta\rho_{ij} \sim \text{Unif}(-1, 1)$ between molecular variables i and j that differ with

respect to interaction range and strength. For illustration purposes, these different potential ranges are shown in Supplementary Figure S1 for $\phi_{ij} = 5$ (solid line), $\phi_{ij} = 10$ (dashed line), and $\phi_{ij} = 20$ (dotted line). We found that SpaCeNet achieves an edge-recovery performance, with respect to both measures AUROC and AUPRC, that is almost as good as in the more controlled setting of the previous simulation studies, in particular when the total number of cells $n \cdot S$ is increased.

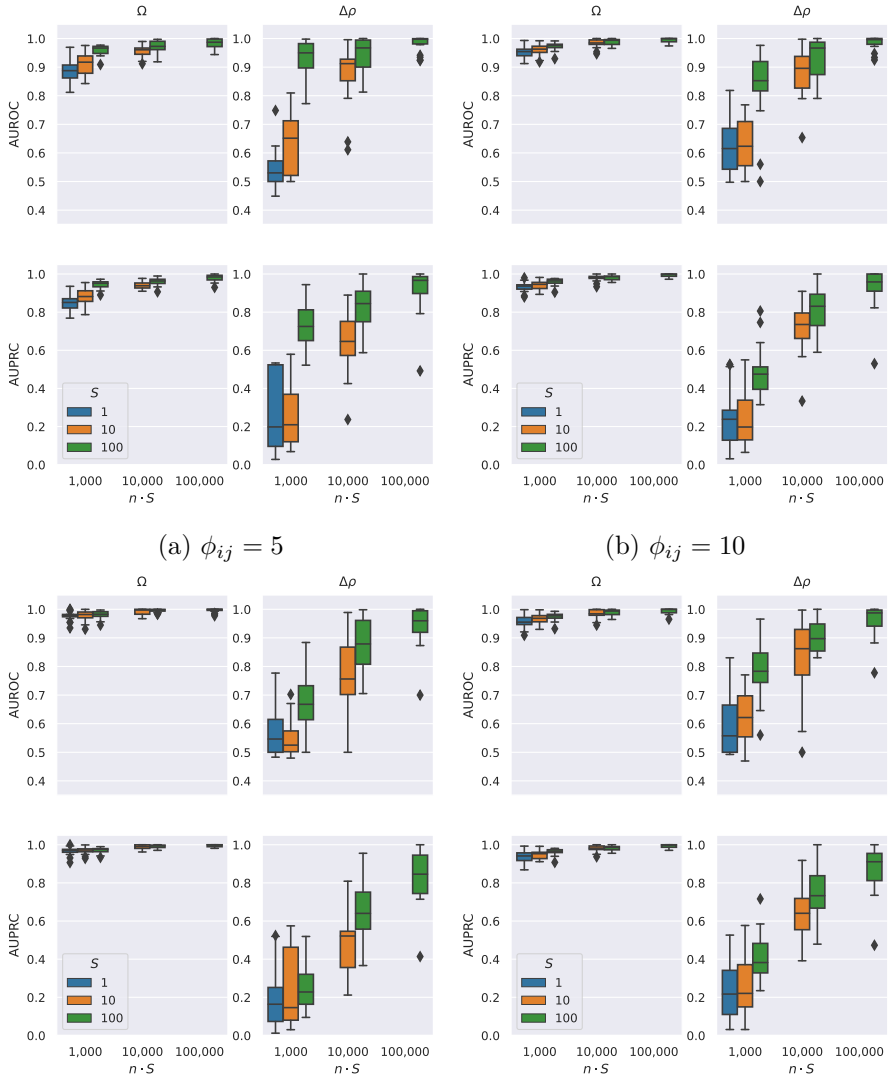
Results of the simulation studies are summarized in Supplementary Table 1. An exemplary SpaCeNet model obtained in this study together with its ground truth is shown in Supplementary Figure S2 for two different choices of the model regularization parameter β.

2.3. Spatial gene expression associations in the mouse visual cortex

We use SpaCeNet to analyze spatial transcriptomics data from the mouse visual cortex generated by STARmap [Wang et al., 2018]. STARMap labels and amplifies cellular RNAs. These amplicons are then transferred to a hydrogel while lipids and proteins are removed. The hydrogel is optically transparent and can be sequentially imaged through multiple cycles with a low probability of errors and miscodings. STARmap measurements of the mouse visual cortex were downloaded from https://www.starmapresources.org/data. The data consist of measurements for 28 RNAs in ~30,000 cells together with their respective location in a $0.1 \times 1.4 \times 1.7$ mm tissue section.

We used SpaCeNet to estimate a global, intracellular network (the precision matrix $\mathbf{\Omega}$) and a network of spatial molecular interactions (the spatial interaction parameters $\mathbf{\Delta\rho}^{(\cdot)}$), where we set $L = 3$ and split the data into four equally sized batches of which three served for model building and one for model validation and hyper-parameter calibration (Suppl. Fig. S3). We tuned α and $\beta \in [10^{-5}, 10]$ on a 4×4 grid that was refined 6 times (Suppl. Fig. S4) and selected the best set of hyper-parameters based on the highest validation pseudo-log-likelihood. A final model was estimated using the full data.

For the intracellular precision matrix $\mathbf{\Omega}$ we obtained a complete matrix with weights summarized in Supplementary File 1. For the spatial interactions $\mathbf{\Delta\rho}^{(\cdot)}$, SpaCeNet selected a set of 134 out of 406 possible spatial associations (Suppl. Fig. S5, Suppl. File 1). We ranked edges according to $\Delta\rho_{ij} = \sqrt{\sum_l \Delta\rho_{ij}^{(l)^2}}$, for which the greatest edge weight was between RNAs encoded by the genes myelin basic protein (*Mbp*) and FMS-related receptor tyrosine kinase 1 (*Flt1*) with $\Delta\rho_{Mbp,\ Flt1} = 0.328$ and $\mathbf{\Delta\rho}^{(\cdot)}_{Mbp,\ Flt1} = (0.327, -0.026, -0.001)^\top$, where the entries of $\mathbf{\Delta\rho}^{(\cdot)}_{Mbp,\ Flt1}$ correspond to the different orders in the expansion. Negative/positive values of $\Delta\rho_{ij}^{(l)}$ corre-



(a) $\phi_{ij} = 5$

(b) $\phi_{ij} = 10$

(c) $\phi_{ij} = 20$

(d) $\phi_{ij} \sim \text{Unif}(5, 20)$

Figure 3: Edge recovery by SpaCeNet assessed in a simulation study using (a-c) fixed, radially decreasing potentials $\rho(r_{ab}) = \Delta\rho_{ij} \exp(-\phi_{ij} r_{ab})$ for all cell-cell interactions with constant range parameters $\phi_{ij} \in \{5, 10, 20\}$, and (d) flexible potentials $\rho_{ij}(r_{ab}) = \Delta\rho_{ij} \exp(-\phi_{ij} r_{ab})$ with $\phi_{ij} \sim \text{Unif}(5, 20)$ that mediate the interaction between molecular variables i and j. The y-axes give the performance in terms of the area under the receiver operating characteristics curve (AUROC, top row) and in terms of the area under the precision recall curve (AUPRC, bottom row). Left figures correspond to the inner-cellular networks (the intracellular precision matrix $\mathbf{\Omega}$) and the right figures to the extracellular networks (the cell-cell interaction parameters $\mathbf{\Delta\rho}$). The x-axis stratifies the analysis with respect to total cell numbers $n \cdot S$, where n is the number of cells within a measurement and S the number of measurements. Here, $S = 1$ is shown in blue, $S = 10$ in orange, and $S = 100$ in green.

spond to positive/negative spatial associations in analogy to the definition of the precision matrix $\mathbf{\Omega}$.

First, we verified the absolute residuals between the observed data matrix $\mathbf{X}$ and its SpaCeNet reconstruction $\hat{\mathbf{X}}$ using either (a) the intracellular edges only ($\mathbf{\Delta\rho}^{(\cdot)} = \mathbf{0}$), or (b) the full model (see Suppl. Section A). The results corresponding to this analysis are shown in Figure 4a and b, respectively. We observed that the spatial associations improve model building substantially, as illustrated by the band of large residuals in the upper left quadrant for both *Flt1* and *Mbp* (Fig. 4a and d), that were reduced substantially via the spatial interactions (Fig. 4b and e). Figure 4c and f show the corresponding contributions from the spatial interactions $\mathbf{\Delta\rho}^{(\cdot)}$.

The connection between *Mbp* and *Flt1* was interesting for a number of reasons. Myelin basic protein is the second most abundant protein, after proteolipid protein, of the myelin membrane in the central nervous system (CNS), making up approximately 30% of the total protein of myelin [Boggs, 2006] and *Mbp* mutants lack compact myelin in the CNS [Readhead et al., 1990]. Myelin surrounds nerve cell axons to insulate them and to increase the conduction of electric impulses [Bean, 2007]. Demyelinating diseases of the CNS, of which multiple sclerosis (MS) is the most common, are characterized by damaged myelin sheaths [Love, 2006]. *Flt1* encodes a member of the vascular endothelial growth factor receptor (VEGFR) family. VEGFRs mediate diverse cellular communication signals controlling developmental processes, such as neurogenesis or gliogenesis [Wittko-Schneider et al., 2013]. VEGFRs recognize vascular endothelial growth factors (VEGFs), whose expression has been shown to be upregulated in both acute and chronic MS plaques [Proescholdt et al., 2002]. Our SpaCeNet model resolves that *Flt1* levels are negatively associated with *Mbp* levels in neighboring cells. A possible interpretation is that cells expressing *Flt1* accumulate in the spatial vicinity of cells which lack *Mbp*. This would be in line with the observation that VEGF-A (which is recognized by Flt1) promotes migration of oligodendrocyte precursor cells (OPCs) in a concentration-dependent manner [Hayakawa et al., 2011], as shown by anti-Flk-1 (not Flt1) receptor-blocking antibody. Moreover, it was shown in the medulla oblongata of the adult mouse that OPCs contribute to focal remyelination and that VEGF signaling might be required for their proliferation [Hiratsuka et al., 2019]. It has been long known that Oligodendrocytes are the myelinating cells of the CNS derived from OPCs [Pfeiffer et al., 1993], therefore our finding could be a hint that Flt1 mediates signals necessary to guide myelinating cells such as OPCs and mature oligodendrocytes to axons that lack myelin.

It should be noted that SpaCeNet models are associative, not causal. Thus, we can not elucidate whether decreased *Mbp* levels imply increased *Flt1* in neighboring cells or if increased *Flt1* levels imply decreased *Mbp* in neighboring cells. Moreover, since only a selected set of 28 genes was spatially assessed by STARmap in a thick tissue section, and since numerous signal-

ing steps may be involved in establishing this relationship, the underlying mechanisms remain to be fully elucidated.

The spatial association *Flt1-Mbp* was followed in strength by *Ctgf-Gja1*, *Ctgf-Pcp4*, and *Reln-Sst* associations. *Ctgf* encodes connective tissue growth factor (Ctgf), also known as cellular communication network factor 2 (Ccn2), which belongs to the CCN family. CNN proteins are a family of extracellular matrix proteins involved in intercellular signaling [Jun and Lau, 2011]. We obtained $\Delta\rho_{Ctgf,\ Gja1} = 0.103$, $\boldsymbol{\Delta\rho}^{(\cdot)}_{Ctgf,\ Gja1} = (0.1023, -0.0060, -0.0001)^\top$. Ctgf has been reported to facilitate gap junction intercellular communication in chondrocytes through up-regulation of connexin 43 (Gja1) [Wu et al., 2021]. Purkinje Cell Protein 4 (Pcp4) regulates calmodulin activity and might contribute to neuronal differentiation through the activation of calmodulin-dependent kinase signaling pathways [Mouton-Liger et al., 2011]. The *Ctgf-Pcp4* potential was estimated with $\Delta\rho_{Ctgf,\ Pcp4} = 0.083$ obtained from $\boldsymbol{\Delta\rho}^{(\cdot)}_{Ctgf,\ Pcp4} = (0.0831, -0.0012, -0.0004)^\top$; we are not aware of a mechanism that might explain the spatial association observed between *Ctgf* and *Pcp4*. This association was followed by *Reln* (Reelin) - *Sst* (Somatostatin) with parameters $\Delta\rho_{Reln,\ Sst} = 0.083$ and $\boldsymbol{\Delta\rho}^{(\cdot)}_{Reln,\ Sst} = (0.0780, 0.0034, 0.0008)^\top$. The gene *Reln* encodes a secreted extracellular matrix protein that might play a role in cell-cell interactions critical for cell positioning. Somatostatin affects transmission rates of neurons in the CNS and cell proliferation. GABAeric cortical interneurons can be delineated to 95% by the markers Pv, Sst (co-expressed with Reelin), Reelin (without Sst), and Vip [Miyoshi et al., 2010]. The medial ganglionic eminence gives rise to the population of interneurons that co-express Reelin and Somatostatin, while the caudal ganglionic eminence gives rise to interneurons that express Somatostatin but lack Reelin [Miyoshi et al., 2010]. The observed spatial *Reln-Sst* association could be a hint that the distribution of these two populations is spatially well organized and not random.

2.4. Spatial gene expression interaction patterns in the *Drosophila* blastoderm

Single-cell RNA sequencing with single-cell spatial resolution is still in its early days and so far captures only a small proportion of cells in a specimen. For example, sci-Space captures only about 5% [Srivatsan et al., 2021]. For the previous analysis, we used data generated by STARMap [Wang et al., 2018], an in-situ sequencing technology for dense measurements at single-cell resolution, which, however, is limited in throughput. Thus, full transcriptome single-cell resolved spatial data do not yet exist. In contrast, omics readouts together with spatial information (the transcriptome in a given spatially defined region of a specimen) become more and more available. Here, we illustrate that SpaCeNet augments their analysis.

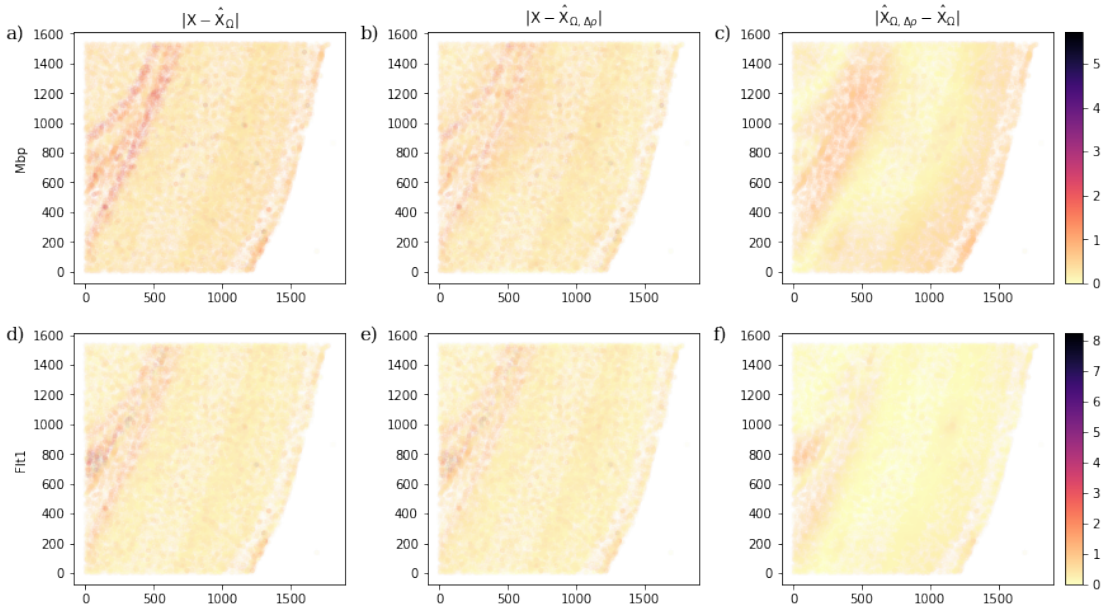
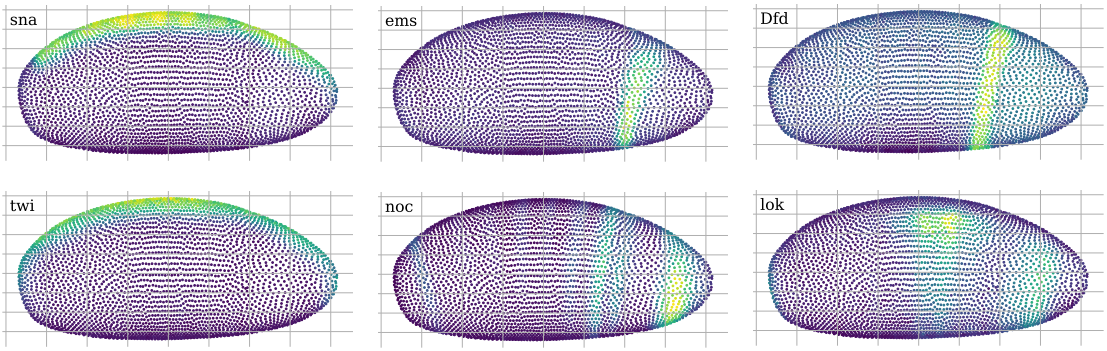


Figure 4: Absolute prediction residuals for spatially associated genes *Mbp* (top row) and *Flt1* (bottom row) in the mouse visual cortex data from Wang et al. [2018]. The left column shows $|\mathbf{X} - \hat{\mathbf{X}}_{\boldsymbol{\Omega}}|$, which includes only the residuals taking into account the intracellular interaction parameters ($\boldsymbol{\Omega}$) and neglecting the spatial interactions ($\boldsymbol{\Delta\rho}^{(\cdot)} = 0$). The middle column displays the residuals if both intra- and intercellular interactions are considered, $|\mathbf{X} - \hat{\mathbf{X}}_{\boldsymbol{\Omega},\boldsymbol{\Delta\rho}}|$. The right column shows the contributions from spatial interactions only, $|\hat{\mathbf{X}}_{\boldsymbol{\Omega},\boldsymbol{\Delta\rho}} - \hat{\mathbf{X}}_{\boldsymbol{\Omega}}|$.

The Berkeley Drosophila Transcription Network Project used a registration technique that uses image-based data from hundreds of *Drosophila* blastoderm embryos, each co-stained for a reference gene and one gene out of a preselected gene set, to generate a virtual *Drosophila* embryo [Fowlkes et al., 2008]. We retrieved these virtual embryo data consisting of 84 genes whose expression levels were measured at 3039 embryonic locations[1]. We then used SpaCeNet to estimate the intracellular network (the precision matrix $\boldsymbol{\Omega}$) and the network of spatial molecular interactions ($\boldsymbol{\Delta\rho}^{(\cdot)}$) for $L = 3$. We performed a hyper-parameter grid search, where we trained the model on 70% of the data and validated it on the remaining 30% (Suppl. Fig. S6 and S7). The best set of hyper-parameters was selected based on the highest validation pseudo-log-likelihood and was used to fit a final model on the full data (Suppl. File 2). The intracellular network $\boldsymbol{\Omega}$ is a full matrix and so there is a rich dependency structure among variables not related to their spatial context. The spatial-interaction network, $\boldsymbol{\Delta\rho}^{(\cdot)}$, in contrast, is sparse with

[1] gene expression: http://bimsbstatic.mdc-berlin.de/rajewsky/DVEX/bdtnp.txt.gz
3D coordinates: http://bimsbstatic.mdc-berlin.de/rajewsky/DVEX/geometry.txt.gz



(a) $\Delta\rho^{(1)}_{twi,\ sna} = -0.036$ (b) $\Delta\rho^{(1)}_{ems,\ noc} = 0.030$ (c) $\Delta\rho^{(1)}_{Dfd,\ lok} = 0.021$

Figure 5: PCA plot of Drosophila observation coordinates with expression levels of genes highlighted. (a) Genes *twi* and *sna* show high expression levels in the same area. (b & c) Gene pairs (*ems*, *noc*) and (*Dfd*, *lok*) are activated in adjacent but different areas.

238 out of 3570 possible interactions (Suppl. Fig. S8). Spatial molecular interactions improved goodness-of-fit on validation data and thus also the generalizability of the model, which highlights the need to include spatial interactions.

Figure 5 contrasts the expression of the gene pairs *twi*-*sna*, *ems*-*noc* and *Dfd*-*lok*, that showed the highest spatial association in the SpaCeNet analysis with $\Delta\rho_{twi,\ \mathrm{sna}} = 0.0368$, $\Delta\rho_{ems,\ noc} = 0.0306$, and $\Delta\rho_{Dfd,\ lok} = 0.0211$. We observed *twi* and *sna* to be active in the same spatial regions, which is consistent with both the positive spatial association suggested by SpaCeNet in leading order ($\Delta\rho^{(1)}_{twi,\ \mathrm{sna}} = -0.0361$) and with the joint activation of *twi* and *sna* in the differentiation of the *Drosophila* mesoderm in localized (ventral) regions of early embryos [Ip et al., 1992]. In contrast, the genes *ems* and *noc* (for which SpaCeNet estimated a negative spatial association at leading order ($\Delta\rho^{(1)}_{ems,\ noc} = 0.0301$)) are active in adjacent but different areas of the *Drosophila* embryo. A similar observation can be made for the genes *Dfd* and *lok*, for which SpaCeNet also estimated a negative leading-order association ($\Delta\rho^{(1)}_{Dfd,\ lok} = 0.0210$).

2.5. Inference of spatial gene expression patterns in the *Drosophila* blastoderm

SpaCeNet infers a joint density function describing spatially distributed, potentially high-dimensional molecular features. Thus, it can be also utilized as an inferential tool, predicting the gene expression of a cell given its cellular context. For illustration purposes, consider the gene *Kr* that encodes the Krüppel protein, a transcriptional repressor expressed in the center of the

embryo during the cellular blastoderm stage [Licht et al., 1990]. First, we predicted *Kr* expression in a leave-one-position-out approach that provides levels of *Kr* based on a cell's environment. The ground truth is shown in Figure 6a and the corresponding predictions in Figure 6b. We found that the SpaCeNet model so trained can predict each cell's individual expression, given its location in space (mean squared error (MSE) $= 4.36 \cdot 10^{-3}$). However, slight deviations were observed in the highlighted region (red arrow). Next, we tested whether this discrepancy could be resolved by specifying *a priori* the expression levels of the remaining genes at the position of interest (Fig. 6c). This reduced the MSE to $2.88 \cdot 10^{-3}$, and the highlighted region better agrees with the ground truth, as seen from comparison of Figure 6a to c.

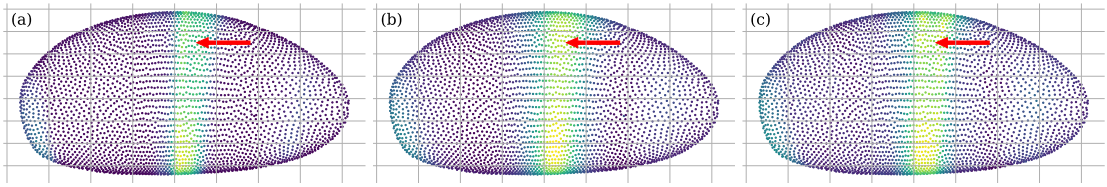


Figure 6: Ground truth and predictions of *Kr* gene expression in Drosophila based on different densities: (a) Observed expression. (b) Prediction of $\mathbf{x}^a$ based on density $f(\mathbf{x}^a|\mathbf{R}, \mathbf{X}^{\setminus a})$, meaning that the expression $\mathbf{x}^a$ is predicted using the expression levels of all other cells $\mathbf{x}^b$ with $b \neq a$. (c) Prediction of x_j^a based on density $f(x_j^a|\mathbf{R}, \mathbf{X}^{\setminus a}, \mathbf{x}_{\setminus j}^a)$, meaning that the expression $\mathbf{x}_j^a$ is predicted using the expression levels of all other cells $\mathbf{x}^b$ with $b \neq a$ and the expression levels of cell a except j. The red arrow highlights an area where the predictions (b) and (c) differ most.

3. Discussion

SpaCeNet is a network inference method that determines intracellular correlation networks and cell-cell associations from spatial molecular data. SpaCeNet is based on probabilistic graphical modeling and extends the concept of conditional independence (CI) by spatial information through estimating spatial conditional independencies (SCI). These intracellular and intercellular SCI relationships encode information about how molecular variables affect each other across space. We verified SpaCeNet in comprehensive simulation studies and demonstrated the information that SpaCeNet can extract from spatial transcriptomics data in two example data sets: an expression map of the mouse visual cortex and a virtual RNA map of the *Drosophila* blastoderm. The analysis of the mouse visual cortex allowed us to generate hypotheses about the spatial organization of cell populations, such as *Flt1-*

mediated signals which could be involved in the recruitment of myelinating cells towards axons which lack myelin basic protein (*Mbp*), and the analysis of the *Drosophila* blastoderm showed that SpaCeNet can also yield insights if data do not have cellular resolution. In the latter case, SpaCeNet resolved spatial association patterns between molecular variables, as observed between the gene-pairs *twi-sna* and *ems-noc*.

Modern spatial transcriptomics techniques are capable of measuring RNA transcript levels with single-cell resolution in a three-dimensional space. Many molecular processes that take place within and between cells, such as translation to protein, possible post-translational modifications, and subsequent signaling cascades comprising secretion of molecules and their recognition by other cells, are not directly measured by these technologies. Although laboratory methods under development, such as single-cell proteomics and metabolomics [Marx, 2019, Duncan et al., 2019] or single-cell epigenetic or chromatin confirmation measurements may one day bridge this gap, at present one can only infer such associations with appropriate analytical methods. SpaCeNet is an important first step in inferring complex, multi-omic, intracellular and intercellular associations networks that takes advantage of high-dimensional omics data and spatial information. The present implementation of SpaCeNet uses single-cell spatial transcriptomics data and so can only infer associations between RNA expression levels within and, more importantly, between different cells – without knowledge about any potentially complex intermediate processes. SpaCeNet is based on a robust Gaussian Graphical Model and partial correlations that can scale to the increasingly large and complex, spatial multi-omics data sets that new laboratory technologies are providing.

SpaCeNet's current implementation is limited to modeling continuous variables. While many omics measurements such as expression are continuous, some single-cell assays, such as mutations or chromatin accessibility can be discrete and so are better modeled as categorical variables. SpaCeNet could be easily adapted to include categorical variables using mixed graphical modeling suggested by [Lee and Hastie, 2015]. As is generally the case for correlation-based methods, SpaCeNet does not identify causal interactions, although lower bounds on causal effects can be derived from observational data [Maathuis et al., 2009], and techniques such as directed acyclic graphs and respective equivalence classes may be adapted to SCI relationships, which we intend to explore in the future.

Nevertheless, SpaCeNet represents an important step forward in the analysis of spatial expression data, allowing us to move from a simple atlas of expression values and cell types to models that capture complex patterns of interactions that allow tissues to function and guide cellular growth, development, and disease processes. As new experimental techniques deliver larger and more complex multi-omics data combined with higher resolution information on the location of individual cells, techniques like SpaCeNet will

become increasingly important for integrating spatial and biological contexts.

4. Methods

4.1. SpaCeNet model

4.1.1. Full probability density

Let $\mathbf{X} = [\mathbf{x}^1, \dots, \mathbf{x}^n]^\top \in \mathbb{R}^{n \times p}$ be a data matrix of gene expression levels with p genes in its columns and n cells in its rows. All cell-cell distances are collected in a matrix $\mathbf{R} = (r_{ab}) \in \mathbb{R}^{n \times n}$, where $r_{ab} = r_{ba}$ corresponds to the Euclidean distance between cells a and b. SpaCeNet assumes that data follow an np-dimensional multivariate normal distribution with precision matrix $\mathbf{\Lambda}$, where the entries in $\mathbf{\Lambda}$ capture both intracellular and intercellular spatial conditional independence (SCI) relations.[2]

The *full probability density* function of SpaCeNet is given by

$$f_{\text{full}}(\mathbf{X}|\mathbf{R}) = \sqrt{\frac{\det \mathbf{\Lambda}}{(2\pi)^{np}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\xi} - \mathbf{m})^\top \mathbf{\Lambda}(\boldsymbol{\xi} - \mathbf{m})\right\}, \tag{1}$$

where we stack the individual cells' profiles $\mathbf{x}^a \in \mathbb{R}^p, a = 1, \dots, n$ vertically in $\boldsymbol{\xi} = \text{vec}(\mathbf{X}^\top)$ and use a global, location-agnostic mean vector $\boldsymbol{\mu}$ for all $\mathbf{x}^a$ such that $\mathbf{m} = \mathbf{1}_n \otimes \boldsymbol{\mu} = (\mu_1, \dots, \mu_p, \mu_1, \dots, \mu_p, \dots)^\top$. We decompose the precision matrix $\mathbf{\Lambda} \in \mathbb{R}^{np \times np}$ into

$$\mathbf{\Lambda} = \mathbf{\Lambda}_{\text{within}} + \mathbf{\Lambda}_{\text{between}} \tag{2}$$

with a matrix for within-cell associations

$$\mathbf{\Lambda}_{\text{within}} = \mathbb{I}_n \otimes \mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega} & \mathbf{0}_{p\times p} & \cdots & \mathbf{0}_{p\times p} \\ \mathbf{0}_{p\times p} & \mathbf{\Omega} & & \vdots \\ \vdots & & \ddots & \mathbf{0}_{p\times p} \\ \mathbf{0}_{p\times p} & \cdots & \mathbf{0}_{p\times p} & \mathbf{\Omega} \end{pmatrix} \tag{3}$$

and a matrix for between-cell associations

$$\mathbf{\Lambda}_{\text{between}} = \begin{pmatrix} \mathbf{0}_{p\times p} & \mathbf{\Lambda}^{12}_{\text{between}} & \cdots & \mathbf{\Lambda}^{1n}_{\text{between}} \\ \mathbf{\Lambda}^{21}_{\text{between}} & \mathbf{0}_{p\times p} & & \vdots \\ \vdots & & \ddots & \mathbf{\Lambda}^{(n-1)n}_{\text{between}} \\ \mathbf{\Lambda}^{n1}_{\text{between}} & \cdots & \mathbf{\Lambda}^{n(n-1)}_{\text{between}} & \mathbf{0}_{p\times p} \end{pmatrix}. \tag{4}$$

Due to the parametrization of $\mathbf{\Lambda}_{\text{within}}$ in Eq. (3), the same intracellular precision matrix $\mathbf{\Omega} = (\omega_{ij}) \in \mathbb{R}^{p \times p}$ accounts for the within-cell (intracellular)

[2] Note that space-agnostic conditional independence (CI) relations for gene expression levels of a single cell a can be obtained from our model by marginalizing over all variables of all other cells $b \neq a$.

associations for each cell. Note that the conventional (space-agnostic) Gaussian Graphical Model (GGM) is contained in our SpaCeNet model as the special case $\mathbf{\Lambda}_{\text{between}} = \mathbf{0}_{np\times np}$.

We further assume that the between-cell association of gene i in any cell a with gene j in any other cell b is described by an arbitrary function of the cells' Euclidean distance r_{ab}. We denote this radial cell-cell interaction potential by $\rho_{ij}(r_{ab})$ and write $\mathbf{\Lambda}^{ab}_{\text{between}} = \big(\rho_{ij}(r_{ab})\big) \in \mathbb{R}^{p\times p}$, which we require to be symmetric with respect to both $i \leftrightarrow j$ and $a \leftrightarrow b$. The $(np\times np)$-dimensional precision matrix $\mathbf{\Lambda}$ is required to be positive definite for f_{full} to be a valid probability density.

4.1.2. Cell-cell interaction potentials

From the definition of the probability density function, Eq. (1), we see that

$$X_i^a \perp X_j^b | \{\text{rest}\} \quad \text{with} \quad a \neq b \quad \Leftrightarrow \quad \rho_{ij}(r_{ab}) = 0, \tag{5}$$

where the set "rest" refers to all variables of cell a except X_i^a, all variables of cell b except X_j^b, and all variables of all other cells. Thus, all intercellular SCI relations are encoded in the set of $p(p+1)/2$ independent functions $\rho_{ij}(r) = \rho_{ji}(r)$. To ensure that cells that are infinitely separated do not interact, we require $\rho_{ij}(r) = 0$ for $r \to \infty$. To approximate the potential $\rho_{ij}(r)$, we use a power-series in $(1 - e^{-r/r_0})\frac{r_0}{r}$,

$$\rho^*_{ij}(r) = \sum_{l=1}^{L} \Delta\rho_{ij}^{(l)} \left(1 - e^{-r/r_0}\right)^l \left(\frac{r_0}{r}\right)^l . \tag{6}$$

Let l be the order in the series expansion and $\mathbf{\Delta\rho}^{(l)} = (\Delta\rho_{ij}^{(l)})$ be the corresponding coefficient matrices, which we require to be symmetric, $\Delta\rho_{ij}^{(l)} = \Delta\rho_{ji}^{(l)}$. This yields:

$$\mathbf{\Lambda}^*_{\text{between}} = \sum_{l=1}^{L} \mathbf{\Theta}^{(l)} \otimes \mathbf{\Delta\rho}^{(l)} , \tag{7}$$

with

$$\Theta_{ab}^{(l)} = \Theta_{ba}^{(l)} = \begin{cases} 0 & \text{for } a = b, \\ (1 - e^{-r_{ab}/r_0})^l \left(\frac{r_0}{r_{ab}}\right)^l & \text{else,} \end{cases} \tag{8}$$

which approximates the between-cell precision matrix $\mathbf{\Lambda}_{\text{between}}$. The coefficients $\Delta\rho_{ij}^{(l)}$ are model parameters which are estimated using a regularized pseudo-log-likelihood, as outlined below. The expansion (6) naturally fulfills $\lim_{r\to\infty} \rho^*_{ij}(r) = 0$ and $\lim_{r\to 0} \rho^*_{ij}(r) = c_{ij}$ with constants $c_{ij} = \sum_{l=1}^{L} \Delta\rho_{ij}^{(l)}$. Note, an expansion in $(1 - e^{-r/r_0})\frac{r_0}{r}$ has the advantage that terms do not

diverge for $r \to 0$, which is in contrast to an expansion in $1/r$. Thus, the factor $(1 - e^{-r/r_0})$ smoothes the divergence and the amount of smoothing is determined by the additional parameter r_0, which was chosen ad-hoc in our analyses to equal the minimal observed distance between two cells. There are reasonable alternative choices for r_0, e.g., the average nearest-neighbor distance provides a length scale that might be more appropriate for larger numbers of cells n.

4.1.3. Pseudo-log-likelihood

The precision matrix $\mathbf{\Lambda}$ is of size $np \times np$, which makes a naive maximum-likelihood-based estimate (using, for example, a gradient descent) intractable for reasonably large p and n. We address this issue by using a pseudo-log-likelihood approach [Besag, 1975, Lee and Hastie, 2015], which is a computationally efficient and consistent estimator formed by products of all the conditional distributions. Let $\mathbf{R}$ denote the pairwise distances between cells, $\mathbf{X}^{\setminus a}$ denote all gene expression levels in all cells except cell a, and $\mathbf{x}^a_{\setminus j}$ denote all gene expression levels in cell a except gene j. We consider the conditional density

$$\begin{aligned} f_{\mathrm{c}}(x^a_j|\mathbf{R}, \mathbf{X}^{\setminus a}, \mathbf{x}^a_{\setminus j}) = \sqrt{\frac{\omega_{jj}}{2\pi}} \exp\Bigg\{ -\frac{1}{2}\omega_{jj}\Bigg(x^a_j - \mu_j \\ + \frac{1}{\omega_{jj}}\left[\sum_{k\neq j}^{p} \omega_{jk}(x^a_k - \mu_k) + \sum_{b=1}^{n}\sum_{l=1}^{L}\sum_{k=1}^{p}(\Theta^{(l)}_{ab}\Delta\rho^{(l)}_{jk})(x^b_k - \mu_k)\right]\Bigg)^2\Bigg\}, \end{aligned} \tag{9}$$

obtained from the full density given in Eq. (1). This yields the pseudo-log-likelihood

$$\ell(\mathbf{\Omega}, \mathbf{\Delta\rho}^{(\cdot)}, \boldsymbol{\mu}) = \sum_{a=1}^{n}\sum_{j=1}^{p} \log\left(f_{\mathrm{c}}(x^a_j|\mathbf{R}, \mathbf{X}^{\setminus a}, \mathbf{x}^a_{\setminus j})\right), \tag{10}$$

which has the advantage that no potentially high-dimensional matrices have to be inverted for parameter optimization.

4.1.4. Regularization and implementation

Regularization: Parameter regularization has been repeatedly shown to improve the inference of GGMs using, for example node-wise lasso regression [Meinshausen and Bühlmann, 2006], the graphical lasso [Friedman et al., 2008], or covariance shrinkage [Schäfer and Strimmer, 2005b]. This is particularly the case if the number of variables exceeds or is of the same order of magnitude as the number of measurements [Altenbuchinger et al., 2020].

For independent measurements $s = 1, \ldots, S$, the full, regularized pseudo-log-likelihood-based optimization problem of SpaCeNet is given by

$$\underset{\boldsymbol{\Omega}, \boldsymbol{\Delta\rho}^{(\cdot)}, \boldsymbol{\mu}}{\text{minimize}} \left\{ -\sum_{s=1}^{S} \ell_s(\boldsymbol{\Omega}, \boldsymbol{\Delta\rho}^{(\cdot)}, \boldsymbol{\mu}) + nS \left(\alpha \sum_{i<j} |\omega_{ij}| + \beta \sum_{i \leq j} \sqrt{\sum_{l=1}^{L} \left(\Delta\rho_{ij}^{(l)} \right)^2} \right) \right\} . \tag{11}$$

Eq. (11) penalizes the off-diagonal elements of the within-cell precision matrix $\boldsymbol{\Omega}$ via L_1 regularization [Tibshirani, 1996], where the hyper-parameter α calibrates the strength of the regularization. The between-cell interactions are regularized via group-lasso terms [Yuan and Lin, 2006], where the group contains the coefficients of the spatial interactions at different orders of the series expansion. This regularization has the advantage that sparseness is induced simultaneously across all orders in the expansion, so we induce sparseness in the potentials ρ_{ij} and not just in the different orders of its expansion. The corresponding hyper-parameter β calibrates the strength of the regularization.

Implementation: Eq. (11) is a convex optimization problem with lasso and group-lasso regularization terms, which can be efficiently solved via proximal gradient descent [Parikh and Boyd, 2014]. SpaCeNet standardizes the expression data ad hoc to ensure comparable penalization of the variables. A proximal gradient descent with Nesterov acceleration [Nesterov, 1983, Sutskever et al., 2013] is used to optimize the model parameters. The Nesterov method calculates the gradient at an anticipated but not the actual parameter position to improve convergence. We tracked the loss, see Eq. (11), and once its relative change between two iterations was lower than a user-specified threshold, the optimization was terminated. A python implementation of SpaCeNet is available.[3]

4.2. Simulation studies

Data were simulated from the full probability density (1) with potentials $\rho_{ij}(r_{ab}) = \Delta\rho_{ij} \exp(-\phi_{ij} r_{ab})$, using the following procedure:

1. Initialize empty $p \times p$ matrices for $\boldsymbol{\Omega}$ and $\boldsymbol{\Delta\rho}$.
2. Randomly draw a set of symmetric edges (we chose 5% of all possible edges) for $\boldsymbol{\Delta\rho}$ from $\text{Unif}(-1, 1)$ and draw corresponding $\phi_{ij} = \phi_{ji}$ for given positive ϕ_{min} and ϕ_{max} from $\text{Unif}(\phi_{min}, \phi_{max})$.
3. Randomly draw a set of symmetric edges (we chose 10% of all possible edges) for $\boldsymbol{\Omega}$ from $\text{Unif}(-1, 1)$.
4. Draw $s = 1, \ldots, S$ samples of spatial coordinates, uniformly sampled from 3d space with an average density of 100 cells per volume unit,

[3] https://gitlab.gwdg.de/MedBioinf/NetworkInference/SpaCeNet

each with n cells and calculate the pairwise distances $\mathbf{R}^{(s)}$ for each sample.

5. Based on $\boldsymbol{\Omega}$, $\boldsymbol{\Delta\rho}$, ϕ_{ij} and $\mathbf{R}^{(s)}$ construct $\boldsymbol{\Lambda}^{(s)}$ for all s.
6. Calculate the row-wise sum of the absolute values of all $\boldsymbol{\Lambda}^{(s)}$. For each of the p variables, select the maximum value from the corresponding $n \cdot S$ sums, add a small constant (we chose 10^{-7}) and fill the respective diagonal element of $\boldsymbol{\Omega}$ with it to ensure that Λ is positive definite.
7. Update all $\boldsymbol{\Lambda}^{(s)}$, sample $\boldsymbol{\xi}^{(s)}$ from $\boldsymbol{\Lambda}^{(s)}$ by means of a Cholesky decomposition and add a random $\boldsymbol{\mu}$ if desired.

With this setup, we simulated 24 scenarios with different parameter settings for n, S and ϕ_{ij} with 20 independently seeded replicates of each. All simulations used $p = 20$ variables.

In our studies, ϕ_{ij} was chosen between 5 and 20. The rationale behind this choice is motivated as follows. Given a unit density of η, k cells on average occupy a volume $V_k = k/\eta$, which corresponds to a sphere of radius $r_k = (3k/(4\pi\eta))^{1/3}$. We used r_k as an estimate for the average distance between a cell and its k nearest neighbors. For $\eta = 100$, this yields $r_1 \approx 0.13$, $r_8 = 2r_1 \approx 0.27$, and $r_{27} = 3r_1 \approx 0.40$. The range of the potentials $\exp(-\phi_{ij}r)$ can be quantified by $1/\phi_{ij}$. With the average nearest-neighbor distance r_1 as reference point, $\phi_{ij} = 5$ therefore corresponds to a long-range and $\phi_{ij} = 20$ to a short-range potential.

Setting the expansion order $L = 3$, a grid search was performed on all data sets with 4 values for $\alpha \in [10^{-5}, 10^{-3}, 10^{-1}, 10]$ and $\beta \in [10^{-5}, 10^{-3}, 10^{-1}, 10]$ each. The grid is successively refined 6 times such that about 100 different hyper-parameter combinations are evaluated in total. The best set of hyper-parameters was chosen based on the maximum pseudo-log-likelihood of test data. To this end, the full data set was split 70:30 into a training and a test set. If more than one sample was available ($S > 1$), the split was performed between different samples and otherwise ($S = 1$) all observations in 30 % of the spatial volume were used for testing. We initialized the optimization with a step size of 10^{-6}. If overflows were encountered, we successively reduced the step size by a factor of 10. The convergence threshold for the proximal gradient algorithm was set to 10^{-5} and training was terminated after a maximum of 3,088 optimization steps. The AUROC and AUPRC based evaluation methods rely on a threshold-based classification of estimated parameters into positives and negatives. When considering the spatial association parameters we classify an association between two variables i and j to be positive if $|\Delta\rho_{ij}^{(l)}|$ is greater than the threshold for at least one considered order l in the expansion.

Reconstruction of interaction potentials Data for the reconstruction of the interaction potentials (Figure 2) was generated in line with previous

procedure, but using $p = 5$ and only a single spatial edge connecting two of the variables with the potentials given in the caption of Figure 2a to d, respectively. We simulated data for $n = 10$ and $S = 1000$, and the potentials of (6) were fitted with $r_0 = 0.1$.

5. Acknowledgement

The work of HUZ and MA was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant no. 01ZX1912A and 01ZX1912C). NL and MA were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 420069742. JQ and MBG were supported by grants from the National Cancer Institute (U24CA231846 and R35CA220523), JQ was additionally supported by the National Human Genome Research Institute (R01HG011393) of the US National Institutes of Health.

References

Michael Altenbuchinger, Antoine Weihs, John Quackenbush, Hans Jörgen Grabe, and Helena U Zacharias. Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194418, 2020.

Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell reports*, 29(1):202–211, 2019.

Bruce P Bean. The action potential in mammalian central neurons. *Nature Reviews Neuroscience*, 8(6):451–465, 2007.

Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics*, 21(1):482, July 2020.

Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

JM Boggs. Myelin basic protein: a multifunctional protein. *Cellular and Molecular Life Sciences CMLS*, 63(17):1945–1961, 2006.

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015. doi: 10.1126/science.aaa6090. URL `https://www.science.org/doi/abs/10.1126/science.aaa6090`.

Kyle D Duncan, Jonas Fyrestam, and Ingela Lanekoff. Advances in mass spectrometry based single-cell metabolomics. *Analyst*, 144(3):782–793, 2019.

Charless C Fowlkes, Cris L Luengo Hendriks, Soile VE Keränen, Gunther H Weber, Oliver Rübel, Min-Yu Huang, Sohail Chatoor, Angela H DePace, Lisa Simirenko, Clara Henriquez, et al. A quantitative spatiotemporal atlas of gene expression in the drosophila blastoderm. *Cell*, 133(2):364–374, 2008.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Mahsa Ghanbari, Julia Lasserre, and Martin Vingron. The distance precision matrix: computing networks from non-linear relationships. *Bioinformatics*, 35(6):1009–1017, 2019.

Kazuhide Hayakawa, Loc-Duyen D Pham, Angel T Som, Brian J Lee, Shuzhen Guo, Eng H Lo, and Ken Arai. Vascular endothelial growth factor regulates the migration of oligodendrocyte precursor cells. *Journal of Neuroscience*, 31(29):10666–10670, 2011.

Daishi Hiratsuka, Erkin Kurganov, Eriko Furube, Mitsuhiro Morita, and Seiji Miyata. Vegf-and pdgf-dependent proliferation of oligodendrocyte progenitor cells in the medulla oblongata after lpc-induced focal demyelination. *Journal of Neuroimmunology*, 332:176–186, 2019.

Y Tony Ip, Ronald E Park, David Kosman, Karina Yazdanbakhsh, and Michael Levine. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the drosophila embryo. *Genes & development*, 6(8):1518–1530, 1992.

Joon-Il Jun and Lester F Lau. Taking aim at the extracellular matrix: Ccn proteins as emerging therapeutic targets. *Nature reviews Drug discovery*, 10(12):945–963, 2011.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

Je Hyuk Lee, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L. Yang, Thomas C. Ferrante, Richard Terry, Sauveur S. F. Jeanty, Chao Li, Ryoji Amamoto, Derek T. Peters, Brian M. Turczyk, Adam H. Marblestone, Samuel A. Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M. Church. Highly multiplexed subcellular rna sequencing in situ. *Science*, 343(6177):1360–1363, 2014. doi: 10.1126/science.1250212. URL `https://www.science.org/doi/abs/10.1126/science.1250212`.

Jonathan D Licht, Martha J Grossel, James Figge, and Ulla M Hansen. Drosophila krüppel protein is a transcriptional represser. *Nature*, 346 (6279):76–79, 1990.

Seth Love. Demyelinating diseases. *Journal of clinical pathology*, 59(11): 1151–1159, 2006.

Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

Vivien Marx. A dream of single-cell proteomics. *Nature Methods*, 16(9): 809–812, 2019.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

Goichi Miyoshi, Jens Hjerling-Leffler, Theofanis Karayannis, Vitor H Sousa, Simon JB Butt, James Battiste, Jane E Johnson, Robert P Machold, and Gord Fishell. Genetic fate mapping reveals that the caudal ganglionic eminence produces a large and diverse population of superficial cortical interneurons. *Journal of Neuroscience*, 30(5):1582–1594, 2010.

François Mouton-Liger, Sophie Thomas, Revital Rattenbach, Laetitia Magnol, Vanessa Larigaldie, Aurelie Ledru, Yann Herault, Catherine Verney,

and Nicole Créau. Pcp4 (pep19) overexpression induces premature neuronal differentiation associated with ca2+/calmodulin-dependent kinase ii-δ activation in mouse models of down syndrome. *Journal of Comparative Neurology*, 519(14):2779–2802, 2011.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). *Dokl. akad. nauk Sssr*, 269:543–547, 1983.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Steve E Pfeiffer, Art E Warrington, and Rashmi Bansal. The oligodendrocyte and its many cellular processes. *Trends in cell biology*, 3(6):191–197, 1993.

Martin A Proescholdt, Steven Jacobson, Nancy Tresser, Edward H Oldfield, and Marsha J Merrill. Vascular endothelial growth factor is expressed in multiple sclerosis plaques and can induce inflammatory lesions in experimental allergic encephalomyelitis rats. *Journal of Neuropathology & Experimental Neurology*, 61(10):914–925, 2002.

Carol Readhead, Naoki Takasashi, H David Shine, Raul Saavedra, Richard Sidman, and Leroy Hood. Role of myelin basic protein in the formation of central nervous system myelin. *Annals of the new York Academy of Sciences*, 605(1):280–285, 1990.

Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363 (6434):1463–1467, 2019.

Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005a.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005b.

Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2016.10.001. URL `https://www.sciencedirect.com/science/article/pii/S0896627316307024`.

Sanjay R. Srivatsan, Mary C. Regier, Eliza Barkan, Jennifer M. Franks, Jonathan S. Packer, Parker Grosjean, Madeleine Duran, Sarah Saxton, Jon J Ladd, Malte Spielmann, Carlos Lois, Paul D. Lampe, Jay Shendure, Kelly R. Stevens, and Cole Trapnell. Embryo-scale, single-cell spatial transcriptomics. *Science*, 373(6550):111–117, 2021. doi: 10.1126/science.abb9536. URL `https://www.science.org/doi/abs/10.1126/science.abb9536`.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (6294):78–82, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. *International conference on machine learning*, pages 1139–1147, 2013.

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018.

Deborah Weighill, Rebekka Burkholz, Marouen Ben Guebila, Helena U Zacharias, John Quackenbush, and Michael Altenbuchinger. Dragon: determining regulatory associations using graphical models on multi-omic networks. *arXiv preprint arXiv:2104.01690*, 2021.

Ina M Wittko-Schneider, Fabian T Schneider, and Karl H Plate. Brain homeostasis: Vegf receptor 1 and 2—two unequal brothers in mind. *Cellular and Molecular Life Sciences*, 70(10):1705–1725, 2013.

Zuping Wu, Chenchen Zhou, Quan Yuan, Demao Zhang, Jing Xie, and Shujuan Zou. Ctgf facilitates cell-cell communication in chondrocytes via pi3k/akt signalling pathway. *Cell Proliferation*, 54(3):e13001, 2021.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Rong Zhang, Zhao Ren, and Wei Chen. Silggm: An extensive r package for efficient statistical inference in large-scale gene networks. *PLOS Computational Biology*, 14(8):1–14, 08 2018. doi: 10.1371/journal.pcbi.1006369. URL `https://doi.org/10.1371/journal.pcbi.1006369`.

A. Estimates from conditional densities

By construction, we assume a spatially constant mean vector in the full density distribution (1), i.e., every cell is assigned the same expectation $\hat{\mathbf{x}} = \boldsymbol{\mu}$.

However, in the conditional density of Eq. (9), the mean of the normal distribution is shifted to

$$\hat{x}_j^a = \mu_j - \frac{1}{\omega_{jj}} \left[\sum_{k \neq j}^{p} \omega_{jk}(x_k^a - \mu_k) + \sum_{b=1}^{n} \sum_{l=1}^{L} \sum_{k=1}^{p} (\Theta_{ab}^{(l)} \Delta\rho_{jk}^{(l)})(x_k^b - \mu_k) \right] ,$$

which can be used as an estimate for the variable x_j^a provided all other variables are known.

This method was used to calculate the residuals for the mouse visual cortex data (Fig. 4), where $\boldsymbol{\Delta\rho}^{(\cdot)}$ was either set to 0 or the estimated parameters. In this case, information $\mathbf{x}_{\setminus j}^a$ is required, i.e., all other variables of cell a have to be known already.

Similarly to Eq. (9), we determine the common conditional density for all variables of one cell a to be given by

$$\begin{aligned} f(\mathbf{x}^a | \mathbf{R}, \mathbf{X}^{\setminus a}) &= \frac{\sqrt{|\boldsymbol{\Omega}|}}{(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}^a - \boldsymbol{\mu} + \boldsymbol{\Omega}^{-1}\boldsymbol{\rho}^a)^\top \boldsymbol{\Omega} (\mathbf{x}^a - \boldsymbol{\mu} + \boldsymbol{\Omega}^{-1}\boldsymbol{\rho}^a) \right\} \\ &= \frac{\sqrt{|\boldsymbol{\Omega}|}}{(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}^a - \hat{\mathbf{x}}^a)^\top \boldsymbol{\Omega} (\mathbf{x}^a - \hat{\mathbf{x}}^a) \right\} \end{aligned}$$

with

$$\begin{aligned} \hat{\mathbf{x}}^a &= \boldsymbol{\mu} - \boldsymbol{\Omega}^{-1}\boldsymbol{\rho}^a , \\ \rho_j^a &= \sum_{l=1}^{L} \sum_{k=1}^{p} \sum_{b=1}^{n} \Theta_{ab}^{(l)} (x_k^b - \mu_k) \Delta\rho_{jk}^{(l)} . \end{aligned}$$

The mean vector $\hat{\mathbf{x}}^a$ can be used as an estimate for the variables of cell a, provided that all variables of all other cells are known. Note that $\boldsymbol{\rho}^a$ only depends on other cells $b \neq a$ since $\Theta_{aa}^{(l)} = 0$ for all l. A comparison of the results obtained with the two methods is shown in Fig. 6.

B. Tables

Table 1: Results of simulation studies.

				AUROC $\boldsymbol{\Omega}$		AUPRC $\boldsymbol{\Omega}$		AUROC $\boldsymbol{\Delta\rho}$		AUPRC $\boldsymbol{\Delta\rho}$	
				mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\phi_{\min}$	$\phi_{\max}$	$n \cdot S$	S								
5.0	5.0	1000	1	0.89	0.04	0.85	0.04	0.55	0.07	0.28	0.21
			10	0.91	0.04	0.88	0.04	0.63	0.11	0.27	0.17
			100	0.96	0.02	0.94	0.02	0.93	0.07	0.73	0.12
		10000	10	0.95	0.02	0.94	0.02	0.87	0.10	0.64	0.16
			100	0.97	0.02	0.96	0.02	0.94	0.06	0.83	0.11
		100000	100	0.98	0.02	0.98	0.02	0.98	0.03	0.92	0.12
	20.0	1000	1	0.96	0.02	0.94	0.03	0.59	0.10	0.25	0.17
			10	0.97	0.02	0.95	0.02	0.62	0.09	0.26	0.17
			100	0.97	0.01	0.96	0.02	0.78	0.10	0.40	0.12
		10000	10	0.98	0.02	0.98	0.02	0.84	0.13	0.64	0.15
			100	0.99	0.01	0.98	0.01	0.90	0.06	0.75	0.14
		100000	100	0.99	0.01	0.99	0.01	0.96	0.06	0.88	0.13
10.0	10.0	1000	1	0.95	0.02	0.93	0.02	0.62	0.10	0.24	0.15
			10	0.96	0.02	0.94	0.03	0.63	0.09	0.25	0.16
			100	0.97	0.01	0.96	0.02	0.84	0.12	0.48	0.13
		10000	10	0.98	0.01	0.98	0.02	0.88	0.08	0.72	0.13
			100	0.99	0.01	0.98	0.01	0.93	0.07	0.81	0.12
		100000	100	0.99	0.01	0.99	0.01	0.98	0.02	0.92	0.11
20.0	20.0	1000	1	0.98	0.02	0.97	0.02	0.57	0.08	0.20	0.17
			10	0.98	0.02	0.97	0.02	0.55	0.06	0.22	0.20
			100	0.98	0.01	0.97	0.02	0.67	0.10	0.26	0.13
		10000	10	0.99	0.01	0.99	0.01	0.76	0.15	0.49	0.16
			100	1.00	0.01	0.99	0.01	0.88	0.09	0.65	0.14
		100000	100	1.00	0.01	1.00	0.01	0.94	0.07	0.84	0.14

C. Figures

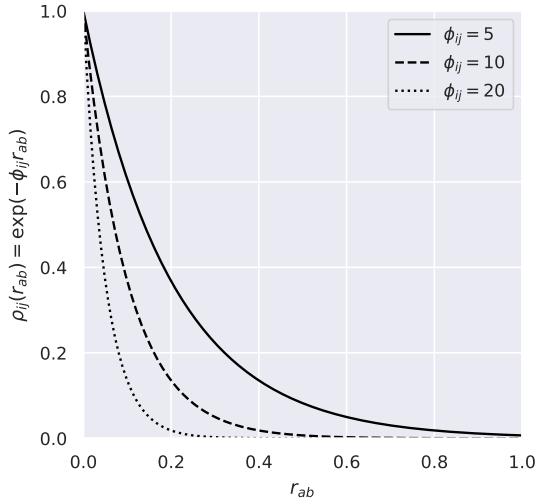


Figure S1: Illustration of different exponential potentials $\rho_{ij}(r_{ab}) = \exp(-\phi_{ij} r_{ab})$ used in the simulation studies, where $\phi_{ij} \sim \text{Unif}(5, 20)$ controls the interaction range. Small values correspond to long range associations ($\phi_{ij} = 5$, solid line), large values to short range associations ($\phi_{ij} = 20$, dotted line) and values in between to medium range associations ($\phi_{ij} = 10$, dashed line).

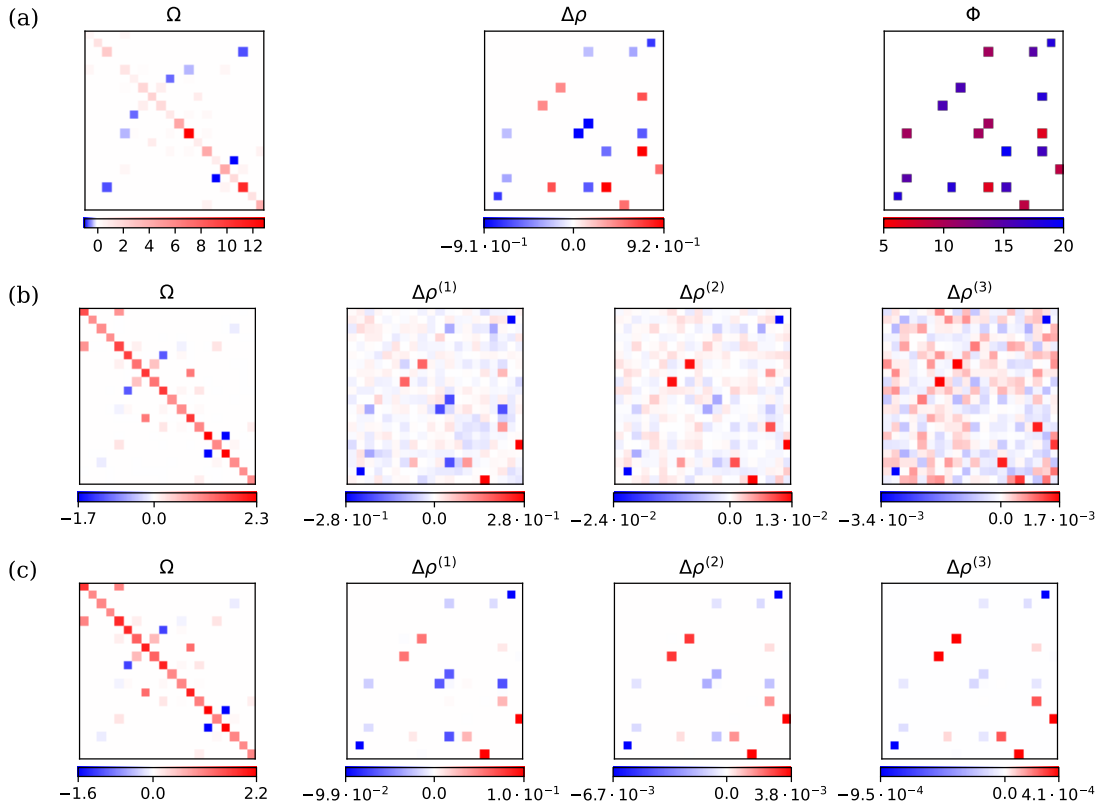


Figure S2: Parameters for a simulated data set with $n = 10^3$, $S = 100$, $\phi_{ij} \in [5, 20]$. (a) True parameters that have been used for sampling. (b) Estimated parameters, where the hyper-parameters were selected based on test-set loss, and (c) estimated parameters, where the hyper-parameters were manually chosen.

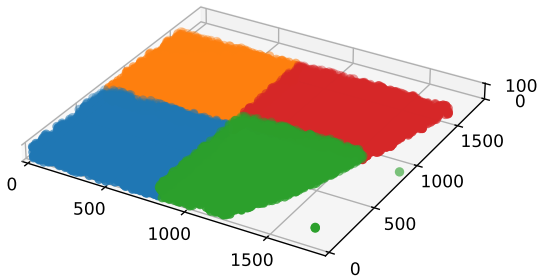


Figure S3: Cellular positions in the mouse visual cortex. Colors indicate the training (blue, orange, green) and validation (red) batches used for hyper-parameter screening and model development.

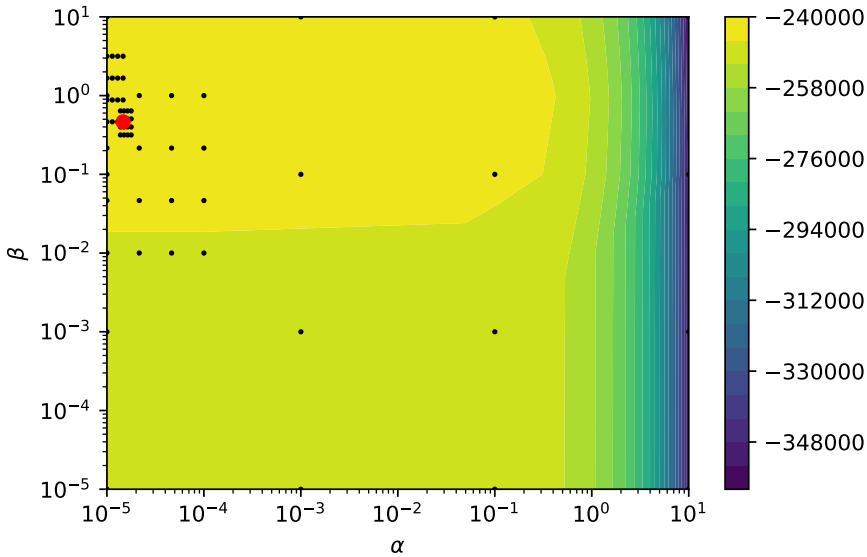


Figure S4: Evaluated hyper-parameter space and corresponding validation pseudo-log-likelihoods colored in blue (low values) to yellow (high values) based on the mouse visual cortex data provided by [Wang et al., 2018]. Black dots are tested hyper-parameters and the red dot corresponds to the optimal set of hyper-parameters in the grid search.

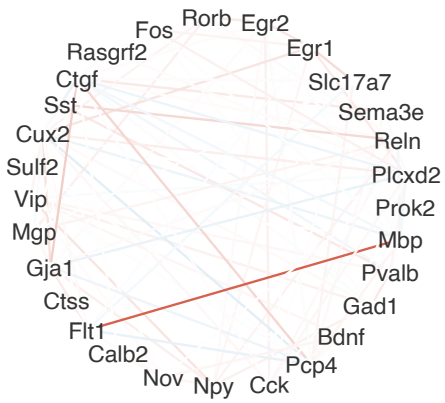


Figure S5: Network of spatial interactions between molecular variables estimated by SpaCeNet using data of the mouse visual cortex provided by [Wang et al., 2018]. Blue edges correspond to positive associations (negative entries of $\boldsymbol{\Delta\rho}^{(\cdot)}$) and red edges to negative associations (positive entries of $\boldsymbol{\Delta\rho}^{(\cdot)}$).

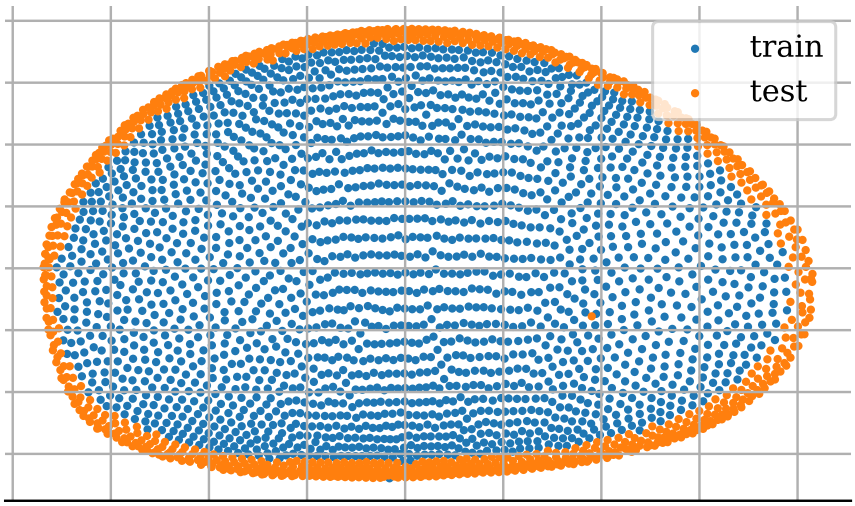


Figure S6: Spatial coordinates (mapped from 3D to 2D via a principal component analysis) of the virtual *Drosophila* embryo [Fowlkes et al., 2008] and the corresponding training/test splitting used for the SpaCeNet hyper-parameter screening shown in blue/orange, respectively.

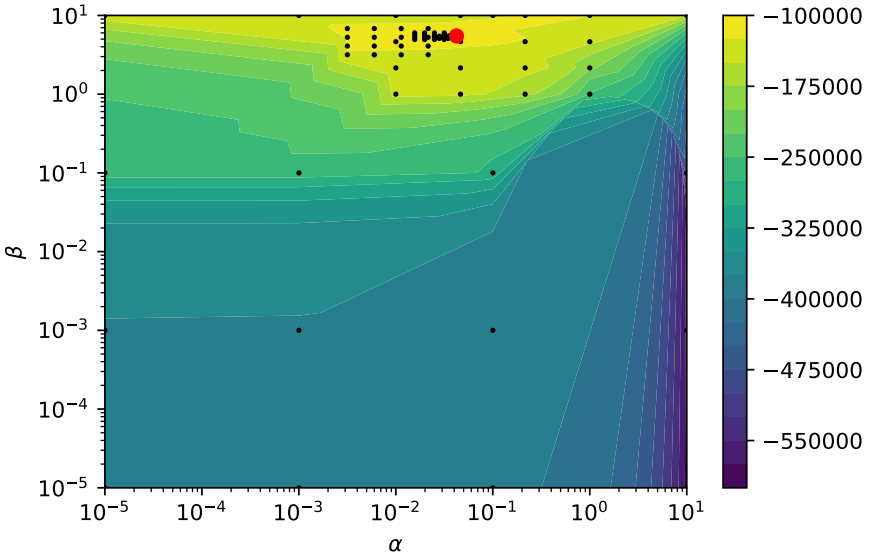


Figure S7: Evaluated hyper-parameter space and corresponding validation pseudo-log-likelihoods colored in blue (low values) to yellow (high values) based on the virtual *Drosophila* embryo data provided by [Fowlkes et al., 2008]. Black dots are tested hyper-parameters and the red dot corresponds to the optimal set of hyper-parameters in the grid search.

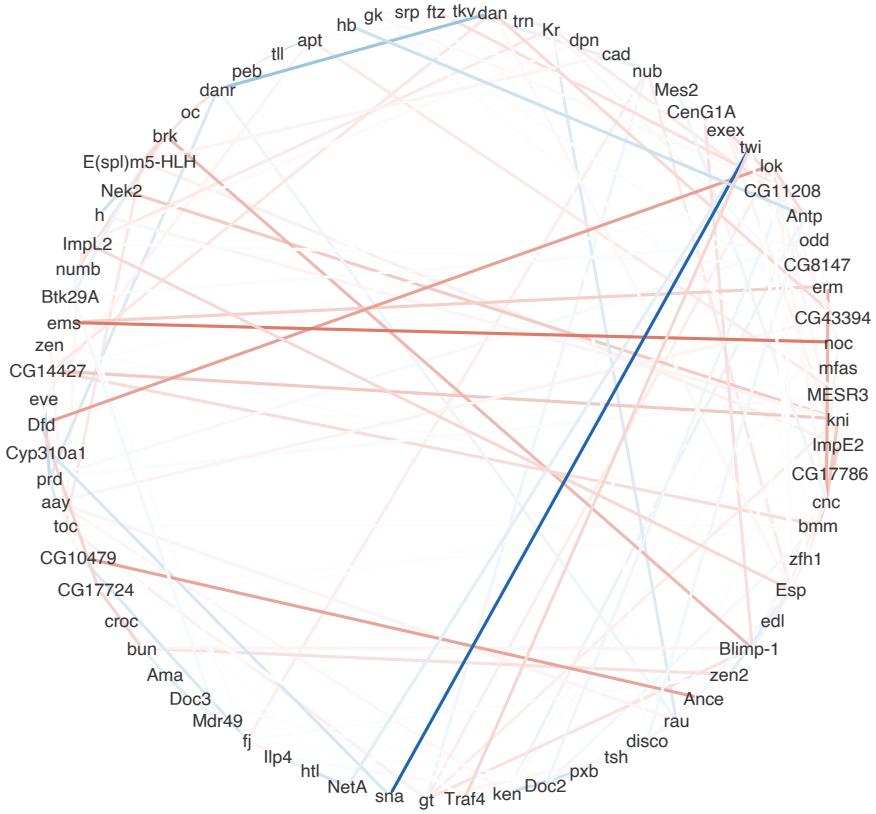


Figure S8: Network of spatial interactions between molecular variables estimated by SpaCeNet using data of the virtual *Drosophila* embryo provided by [Fowlkes et al., 2008]. Blue edges correspond to positive associations (negative entries of $\mathbf{\Delta}\boldsymbol{\rho}^{(\cdot)}$) and red edges to negative associations (positive entries of $\mathbf{\Delta}\boldsymbol{\rho}^{(\cdot)}$).