A Counterfactual Framework for Directional Cell–Cell Interaction Analysis in Spatial Transcriptomics

Humaira Anzum[1*], Veena Kochat[2⋆], Md Ishtyaq Mahmud[1], Jagan Mohan Reddy Dwarampudi[1], Suresh Satpati[2], Pooja Shukla[3], Milind Javle[4], Lawrence Kwong[3], Kunal Rai[2**], and Tania Banerjee[1,5⋆⋆]

[1] Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA
[2] Department of Genomic Medicine and MDACC Epigenomics Therapy Initiative University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3] Department of Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
[4] Department of Gastrointestinal Medical Oncology University of Texas MD Anderson Cancer Center, Houston, TX, USA
[5] Department of Information Science Technology, University of Houston, Houston, TX, USA
tbanerjee@uh.edu, krai@mdanderson.org

Abstract. Understanding neighboring cell influences is central to spatial transcriptomics, yet existing methods rely on correlation or predefined ligand–receptor (LR) pairs and do not test directionality. We introduce a counterfactual, intervention-based framework for inferring directional cell–cell influence that is LR-agnostic and tests sender specificity. A neighborhood graph model predicts receiver cell state from local spatial context. Directional influence is quantified by counterfactually replacing neighbors of a sender type and measuring resulting displacement in predicted receiver state. We define a Counterfactual Directionality Score (CDS) that quantifies directional influence, and compute pairwise CDS by aggregating across receiver cells and test cores sender–receiver pair. Applied to Xenium cholangiocarcinoma tissue microarrays (38 cores), the framework identified reproducible, asymmetric interactions between tumor, immune, and stromal compartments; e.g. Tumor-EMT →Macrophage (CDS: 0.0828) and Fibroblast→Macrophage (CDS: 0.0582). Effects exceeded label-permutation and spatial-shuffle null models ($p < 0.001$, FDR-controlled) and remained stable under core-level bootstrap resampling. Inferred directional strengths correlated strongly with matched LR scores ($r = 0.758$, $p = 0.0027$), supporting biological concordance. Code: `https://github.com/Snigdha022/Counterfactual-Cell-Influence-Spatial-Transcriptomics.git`.

Keywords: Spatial transcriptomics · Cell–cell communication · Counterfactual inference · Tumor microenvironment

⋆ These authors contributed equally.
⋆⋆ Co-corresponding authors

1 Introduction

Spatial transcriptomics measures gene expression while preserving tissue architecture, enabling analysis of cell–cell interactions in situ [5]. In oncology and immunology, understanding how specific cell types influence neighboring cells is central to tumor progression, immune modulation, and therapeutic response [12,8]. However, inferring *directional* cell–cell influence from single-snapshot spatial data remains challenging [11]. Existing computational approaches to cell–cell communication (CCC) are largely correlational. Ligand–receptor (LR)-based methods, including CellPhoneDB [6], CellChat [7], and NicheNet [2], test curated interaction pairs between annotated populations. While interpretable, they depend on prior knowledge and do not explicitly test directionality [9]. Spatial toolkits and transport-based models (e.g., Squidpy [10], Giotto [4], SpaOTsc [4], COMMOT [3]) incorporate spatial structure, and predictive frameworks such as GNNs and transformer-based models (e.g., GITIII [13]) capture complex dependencies. Nevertheless, influence is inferred through predictive attribution rather than explicit intervention.

We instead adopt a counterfactual formulation of directional influence. A neighborhood-conditioned graph model predicts receiver cell state from local spatial context, after which controlled perturbations of sender composition quantify the resulting displacement in predicted receiver state. This enables directional effect quantification through interventional testing of the predictive model, rather than relying on correlation-based association.

Contributions. (1) We introduce a counterfactual, intervention-based framework for analyzing directional cell–cell influence through model interpretability in spatial transcriptomics without predefined LR priors. (2) We define a Counterfactual Directionality Score (CDS) with complementary sender type-swap and within-type shuffle operators to disentangle cell-type identity from intra-type heterogeneity. (3) We establish a rigorous evaluation protocol using matched permutation null models with FDR correction and strict core-level data separation. (4) We enhance interpretability through functional program projection and ligand–receptor concordance analysis.

Additional implementation details are provided in the accompanying extended technical report [1].

2 Proposed Methodology

2.1 Overview

We introduce a counterfactual framework for analyzing directional cell–cell interactions in spatial transcriptomics. A neighborhood-conditioned *Neighbor Influence Model* predicts receiver cell state from local spatial context, after which counterfactual perturbations of sender composition quantify directional influence across independent tissue microarray (TMA) cores. The framework assumes local spatial influence and stable counterfactual perturbations; details are provided in the technical report [1].

2.2 Cell State and Spatial Context Modeling

Let $x_i \in \mathbb{R}^G$ ($G = 480$) denote the expression vector of cell i. To isolate functional state from cell-type identity, we compute a residualized state $s_i = x_i - \mu_{t(i)}$, where $t(i)$ is the cell type and μ_t its mean expression. Residual values are clipped to $[-5, 5]$ and projected to 50 principal components, yielding a denoised state representation $z_i \in \mathbb{R}^{50}$. Each TMA core is treated as an independent biological unit. For every cell i, the $k = 20$ nearest spatial neighbors are identified using Euclidean distance in tissue coordinates. Neighbor states are aggregated via distance-weighted softmax pooling:

$$a_i = \sum_{j \in \mathcal{N}(i)} w_{ij} z_j, \qquad w_{ij} \propto \exp(-d_{ij}),$$

yielding a neighborhood context vector $a_i \in \mathbb{R}^{50}$. The choice of neighborhood size and distance weighting is evaluated through sensitivity analyses (Section 3.5).

We train a *Neighbor Influence Model* to predict receiver state z_i from neighborhood context a_i conditioned on receiver cell type $t(i)$. The model comprises an input batch normalization layer followed by residual hidden layers with layer normalization. Receiver cell type is provided as an index tensor (`recv_type_idx`) and mapped to a learnable embedding via an `nn.Embedding` layer. This embedding is linearly projected and passed through a sigmoid to generate a multiplicative gate, which modulates the hidden representation element-wise, enabling type-specific processing of spatial context. The network is trained using Huber loss ($\delta = 1.0$) between predicted and observed cell states.

Cores are partitioned into training (25), validation (3), and held-out test (10) sets, with counterfactual analyses performed exclusively on test cores. Optimization uses AdamW with early stopping based on validation loss.

2.3 Counterfactual Directionality Score (CDS)

For a sender type S and receiver type R, we quantify directional influence by comparing model predictions under observed versus counterfactual neighborhoods. For each receiver cell i of type R in test cores, we define the vector difference:

$$\Delta_i = f(a_i^{\mathrm{cf}}, t(i)) - f(a_i^{\mathrm{orig}}, t(i)) \in \mathbb{R}^{50},$$

where a_i^{orig} uses the original neighbors and a_i^{cf} is computed from a counterfactual neighborhood. Counterfactual neighborhoods are generated by replacing sender-type neighbor identities while preserving spatial structure via distance binning. For each receiver cell, neighbors are assigned to predefined distance bins and replacement cells are selected from the same core and distance bin. Crucially, only neighbor indices are changed. So, original distances remain unaltered preserving spatial autocorrelation structure. Two complementary counterfactual operators are applied:

Type-swap operator. Neighboring S cells are replaced with non-S cells drawn from the same core, preserving distance bin structure where possible. This tests whether the presence of sender type S directionally influences receiver R. A significant CDS under this null indicates that sender-type identity, rather than spatial proximity alone, drives the observed receiver shift.

Within-type shuffle operator. Neighboring S cells are replaced with *other* cells of the same type S from the same core, preserving sender identity but disrupting specific sender states. This operator tests dependence on sender cell-state heterogeneity. The within-type operator tests whether specific sender functional states are required for the interaction.

For each receiver cell i, we define the CDS as the L1 displacement

$$\mathrm{CDS}_i(S \to R) = \|\Delta_i\|_1.$$

The core-level CDS was calculated by averaging receiver-cell CDS values within each core. Furthermore, the pair-level CDS was computed by averaging core-level CDS values across the 10 held-out test cores. Signed mean deltas were examined separately to characterize the direction of receiver-state shifts (activation vs. suppression). The detailed pseudocode for the counterfactual neighborhood replacement algorithm is presented in Supplementary Algorithm S1 in the technical report [1].

2.4 Program-Level Projection

To enhance biological interpretability, gene-level changes are projected onto predefined functional programs (e.g., EMT, macrophage activation, T-cell exhaustion). For a given functional program P, we compute the program-level delta for cell i by first mapping the counterfactual state difference back to gene expression space using the inverse PCA transformation, and then averaging the resulting gene-level deltas across all genes in the program. Program-level CDS are aggregated identically to gene-level CDS. This provides insight into which biological processes are most affected by a given sender–receiver interaction.

2.5 Statistical Significance

Statistical significance was assessed using three matched null models (each with $B = 200$ permutations) preserving complementary aspects of biological structure: (i) label permutation within cores (tests dependence on cell-type annotation), (ii) distance-bin neighbor shuffling followed by type-swap (tests spatial specificity), and (iii) sender-agnostic replacement (tests sender identity versus generic occupancy). Empirical p-values were computed as $p = \frac{1+\sum_{b=1}^{B} \mathbb{I}(\mathrm{CDS}_{\mathrm{null}}^{(b)} \geq \mathrm{CDS}_{\mathrm{obs}})}{B+1}$, with Benjamini–Hochberg correction applied across all sender–receiver pairs to control the false discovery rate (FDR) at $\alpha = 0.05$. Uncertainty was estimated using block bootstrap over held-out TMA cores (1000 resamples).

2.6 Ligand–Receptor Concordance Validation

For orthogonal biological validation, we assess whether genes driving significant CDS values correspond to known ligand–receptor interactions between sender and receiver types. For each significant pair, we identify genes with largest absolute mean delta and query CellChatDB for interactions where the ligand is expressed in the sender type and the receptor in the receiver type. Only pairs with mean expression > 0.05 in both cell types are retained. The ligand–receptor score is defined as the geometric mean of sender ligand expression and receiver receptor expression. Concordance was quantified by computing the Spearman rank correlation between pair-level CDS values and corresponding ligand–receptor (LR) scores across all evaluated sender–receiver pairs.

3 Experiments

3.1 Dataset and Evaluation Protocol

We evaluated the proposed counterfactual framework on Xenium spatial transcriptomics data from human cholangiocarcinoma tissue micro-arrays (TMA). After doing quality control, we selected 38 independent TMA cores. We then partitioned the cores into 25 training, 3 validation, and 10 held-out test cores. We exclusively computed all reported CDS values, null model analyses, and biological interpretations on the held-out test set. CDS values were first calculated within each core and then aggregated and the uncertainty was estimated using block bootstrap resampling at the core level (B=1000).

3.2 Spatial Model Validation

To verify that our Neighbor Influence (NI) model leverages spatial information rather than learning cell-type-specific expression, we performed a controlled validation on held-out test cores comparing three models: (i) true-NI (authentic spatial neighbors); (ii) shuffled-NI (disrupted neighbor-receiver correspondences); and (iii) a receiver-only MLP baseline (predicts cell state solely from cell type).

Figure 1 shows that true-NI consistently achieves lower prediction error than both baselines. In panel A, it can be seen that the mean squared error (MSE) averaged across 10 held-out cores is lowest for true-NI (0.846), compared to MLP (1.036) and shuffled-NI (0.969). The error bars indicate ± 1 SD. Panel B displays per-core paired comparisons. So for nearly every core, true-NI yields reduced error relative to both baselines, with boxplots confirming downward shift in both median and variance. Paired t-tests confirm significance ($p < 0.001$), which demonstrates that authentic spatial neighborhoods provide measurable predictive benefit.

3.3 Directional Influence of Sender Cell Types

We evaluated directional influence using a sender type-swap counterfactual operator. For each sender–receiver pair, we replaced neighboring sender cells with non-sender cells of a different type from the same core, preserving all other micro-environmental aspects. The resulting shifts in predicted gene expression were aggregated per core as a CDS. Averaging across the 10 held-out test cores yields the pair-level CDS for each sender–receiver pair. Higher CDS indicates stronger directional influence of the sender on the receiver. Several interactions exhibited a statistically significant directional influence. For instance, *Fibroblast* $\rightarrow$ *Macrophage* showed a robust effect (CDS $= 0.0582$, 95% CI $[0.0459, 0.0719]$, FDR < 0.05), while *Tumor-EMT* $\rightarrow$ *Macrophage* (CDS $= 0.0828$, 95% CI $[0.0811, 0.0839]$, FDR < 0.05) represented the largest directional perturbation detected across all pairs.

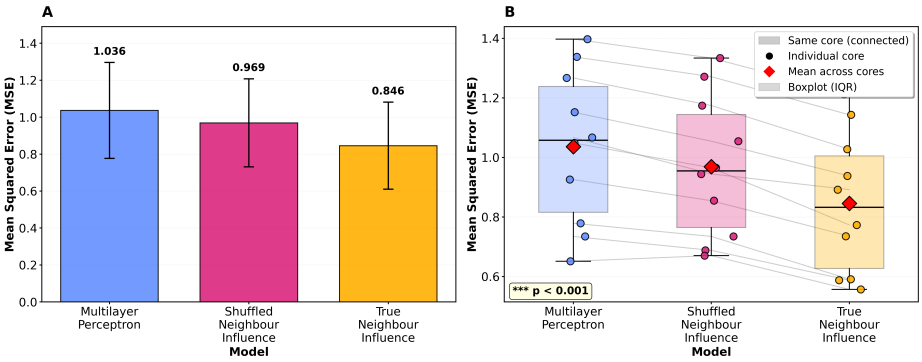

Fig. 1. Performance comparison across held-out test cores. (**A**) Cohort-level MSE with ± 1 SD error bars. (**B**) Per-core paired errors with boxplots and mean markers.

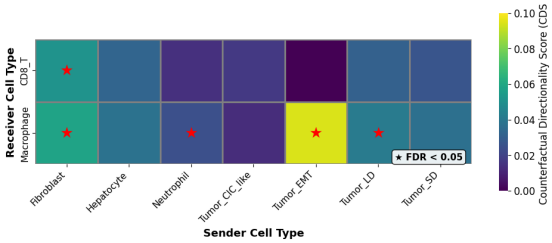

Fig. 2. CDS under type-swap analysis. Red stars denote FDR < 0.05.

3.4 Biological Interpretation of Directional Effects

Gene-level counterfactual deltas were projected onto predefined biological pro grams to assess coherent functional shifts (Figure 3). Tumor-EMT→Macrophage interactions showed significant suppression of Angiogenesis and Macrophage Ac tivation programs, whereas other pairs exhibited weaker or selective effects, indi cating program-specific modulation. Sender–receiver pairs showed coherent bio logical patterns (Fig. 4 of [1]). Tumor-EMT and Tumor-LD reprogrammed macrophage states, while Fibroblast→CD8 T cell interactions suggested stromal suppression of cytotoxic programs. Ligand–receptor concordance analysis using CellChatDB (Figure 4) identified organized interaction structure, with receptors such as CD44, CCR7, and CXCR4 acting as central nodes. This supports interpretation of the directional effects as coordinated signaling modules rather than isolated gene changes Furthermore, we identified the top three up- and down-regulated genes for each significant program-level effect for each pairs. For example, in Fibroblast→Macrophage, Macrophage_Activation was primarily driven by upregulation of CD68, C1QC, and FCGR3A, while EMT showed downregulation of POSTN, COL6A1, and COL6A2. These findings are consistent with the program-level directionality reported above (see Section 3.6). The full lists of up-regulated and down-regulated genes per program are presented in Supplementary Tables S1 and S2, respectively, in the technical report [1].

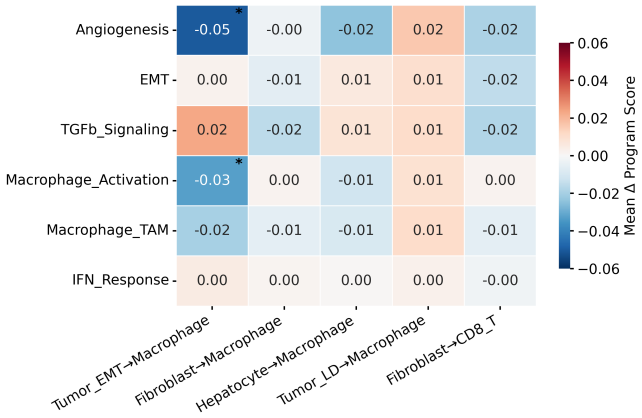

Fig. 3. Program-level mean signed counterfactual deltas (asterisk: FDR < 0.05).

3.5 Robustness Analysis

We evaluated whether the observed effects exceed those expected under randomized spatial structure using matched null models (Fig. 4). Under the type-swap operator (Fig. 4A), the Fibroblast→Macrophage CDS (0.0582) lies in the extreme tail of the label-permutation, neighbor-shuffle, and sender-agnostic null

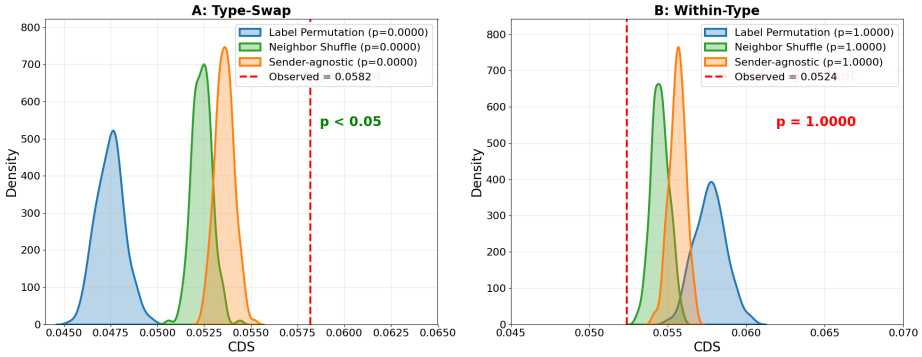

Fig. 4. Null model comparison for Fibroblast→Macrophage. (A) Type-swap randomization. (B) Within-type constrained distribution.

distributions (all $p < 0.001$, FDR-significant), demonstrating robust directional influence that is specific to the true spatial and sender-type configuration. Under the within-type shuffle operator (Fig. 4B), the observed CDS (0.0524) is within all corresponding null distributions (all $p = 1.00$). This indicates that the signal is driven primarily by Fibroblast type identity rather than intra-type state heterogeneity. Bootstrap confidence intervals (reported above) further supported stability across test cores. Per-core CDS estimates are presented in Fig. 6 of the technical report [1].

Sensitivity analyses across neighborhood sizes and distance weighting schemes demonstrated robust CDS estimates. Increasing k produced a monotonic decrease in CDS, motivating the choice of $k = 20$. Complete results are provided in Supplementary Table S3 of the technical report [1].

3.6 Ligand–Receptor Concordance Validation

To evaluate biological concordance, we queried CellChatDB and computed ligand–receptor (LR) scores as the geometric mean of sender ligand and receiver receptor expression (mean expression > 0.05 in both cell types). Across 13 significant sender–receiver pairs, CDS values were strongly correlated with mean LR scores ($r = 0.758$, $p = 0.0027$), indicating alignment between counterfactually inferred directional influence and established signaling pathways. Fibroblast→Macrophage exhibited the highest LR score, consistent with its strong CDS.

Thus, our framework captures biologically meaningful directional influences. But direct numerical comparison with existing methods such as CellChat, NicheNet, or GITIII is not feasible, as they produce fundamentally different outputs (e.g., ligand–receptor co-expression scores or attention weights) which aren't directly comparable to our CDS displacement score. Nevertheless, the strong concordance with CellChatDB ligand–receptor scores provides solid biological validation that our counterfactual framework aligns with established biology.

4 Conclusion

We presented a counterfactual framework for inferring directional cell–cell influence in spatial transcriptomics without predefined ligand–receptor priors. Applied to Xenium cholangiocarcinoma tissue microarrays, the framework identified reproducible and biologically concordant directional interactions between tumor, immune, and stromal compartments. These results establish a statistically grounded approach for model-based directional effect analysis in cell–cell communication. Additional methodological details and supplementary analyses are provided in the technical report [1]. Limitations include dependence on curated gene sets for biological interpretation and evaluation on a single cholangiocarcinoma tissue microarray with 10 held-out test cores. Future work will evaluate the framework across larger multi-patient cohorts. Future work will evaluate the framework on larger multi-patient cohorts and extend it to additional spatial transcriptomics platforms.

Acknowledgments. This project was supported, in part, by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UM1TR004539. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The work was also, in part, supported by STRIDE funding to METI from MD Anderson Cancer Center. We thank the cholangiocarcinoma clinical and pathology group at MDACC for TMA.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anzum, H., Kochat, V., Satpati, S., Mahmud, M.I., Dwarampudi, J.M.R., Shukla, P., Javle, M., Kwong, L., Rai, K., Banerjee, T.: A counterfactual framework for directional cell–cell interaction analysis in spatial transcriptomics: Technical report. `https://raw.githubusercontent.com/taniabanerjee/digital-twin-spatial-transcriptomics/main/papers/MICCAI2026/CDS_MICCAI2026.pdf` (2026), extended technical report
2. Browaeys, R., Saelens, W., Saeys, Y.: Nichenet: modeling intercellular communication by linking ligands to target genes. Nature Methods **17**(2), 159–162 (2020). `https://doi.org/10.1038/s41592-019-0667-5`
3. Cang, Z., Zhao, Y., Almet, A.A., Stabell, A., Ramos, R., Plikus, M.V., Atwood, S.X., Nie, Q.: Screening cell–cell communication in spatial transcriptomics via collective optimal transport. Nature Methods **20**(2), 218–228 (2023). `https://doi.org/10.1038/s41592-022-01728-4`
4. Chen, J.G., Chávez-Fuentes, J.C., O'Brien, M., Xu, J., Ruiz, E.C., Wang, W., Amin, I., Sheridan, J.P., Shin, S.C., Hasyagar, S.V., Sarfraz, I., Guckhool, P., Sistig, A., Jarzabek, V., Yuan, G.C., Dries, R.: Giotto suite: a multiscale and technology-agnostic spatial multiomics analysis ecosystem. Nature Methods **22**(10), 2052–2064 (2025). `https://doi.org/10.1038/s41592-025-02817-w`

5. Chen, T.Y., You, L., Hardillo, J.A.U., Chien, M.P.: Spatial transcriptomic technologies. Cells **12**(16) (2023). https://doi.org/10.3390/cells12162042
6. Efremova, M., Vento-Tormo, M., Teichmann, S.A., Vento-Tormo, R.: Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols **15**(4), 1484–1506 (2020). https://doi.org/10.1038/s41596-020-0292-x
7. Jin, S., Plikus, M.V., Nie, Q.: Cellchat for systematic analysis of cell–cell communication from single-cell transcriptomics. Nature Protocols **20**(1), 180–219 (2025). https://doi.org/10.1038/s41596-024-01045-4
8. Liu, L., Chen, A., Li, Y., Mulder, J., Heyn, H., Xu, X.: Spatiotemporal omics for biology and medicine. Cell **187**(17), 4488–4519 (2024). https://doi.org/10.1016/j.cell.2024.07.040
9. Liu, Z., Sun, D., Wang, C.: Evaluation of cell-cell interaction methods by integrating single-cell rna sequencing data with spatial information. Genome Biology **23**(1), 218 (2022). https://doi.org/10.1186/s13059-022-02783-y
10. Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I., Lotfollahi, M., Richter, S., Theis, F.J.: Squidpy: a scalable framework for spatial omics analysis. Nature Methods **19**(2), 171–178 (2022). https://doi.org/10.1038/s41592-021-01358-2
11. Valihrach, L., Zucha, D., Abaffy, P., Kubista, M.: A practical guide to spatial transcriptomics. Molecular Aspects of Medicine **97**, 101276 (2024). https://doi.org/10.1016/j.mam.2024.101276
12. Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., Haque, A.: An introduction to spatial transcriptomics for biomedical research. Genome Medicine **14**(1), 68 (2022). https://doi.org/10.1186/s13073-022-01075-1
13. Xiao, X., Zhang, L., Zhao, H., Wang, Z.: Inferring spatial single-cell-level interactions through interpreting cell state and niche correlations learned by self-supervised graph transformer. Nature Machine Intelligence **8**(1), 42–58 (2026). https://doi.org/10.1038/s42256-025-01161-0