Single-Cell Cross-Modal Transfer by Adversarial Fine-Tuning of Foundation Models

Joseph Boyd [1] **Matthew Lyon** [1] **Martino Mansoldo** [1] **Christian Hurry** [1] **Finnian Firth** [1]

Abstract

Spatial transcriptomics (ST) is a powerful tool for exploring biological properties dependent on structure, proximity, and interaction in tissue. The methods underpinning ST are developing rapidly but are limited in their ability to profile many thousands of genes at a subcellular scale. Although dissociated from tissue, it is known that the whole-transcriptome readouts of cells in single-cell RNA sequencing (scRNA-seq) retain information about their former *in situ* neighbourhoods, motivating computational methods to recover it. While paired ST and scRNA-seq datasets are scarce, each modality in its own right is abundantly available. We therefore propose to perform cross-modal translation between unpaired ST and scRNA-seq data. In this work we show that a single-cell foundation model can perform this translation via adversarial fine-tuning. We demonstrate that our method performs favourably against methods built for multi-omics translation.

1. Introduction

Methods for spatial transcriptomics (ST) provide spatially resolved readouts of gene expression in biological samples. In particular, fluorescence *in situ* hybridisation (FISH) methods such as MERFISH (Chen et al., 2015) and Xenium (Janesick et al., 2023) provide subcellular resolution to spatial biology, but are often limited in their coverage of the transcriptome. By contrast, single-cell RNA sequencing (scRNA-seq) modalities provide whole-transcriptome readouts, but for cells that have been dissociated from their original tissue context. Due to these differing strengths, ST and scRNA-seq are currently seen as complementary modalities. However, capturing both modalities can be prohibitive in terms of cost and sample availability. Therefore, there is a need for computational methods that can translate between these modalities, allowing one to infer spatial context from scRNA-seq data alone. Furthermore, such a translation may provide insight into the relationship between these modalities, and the different snapshots of biological state they encode.

The aim of this work is to learn to translate, on the level of a single cell, scRNA-seq gene expression to spatially bulked expression deriving from ST. The challenge is to learn to perform this translation from fully unpaired data, and the key insight is to exploit the fact that both modalities manifest as count data. We present single-cell cross-modal (SCXM), an adversarial fine-tuning approach for single-cell foundation models aimed at modality translation from scRNA-seq to ST context features (Figure 1). In so doing, we demonstrate that the gene expression of dissociated cells retains clues as to its former spatial context (Armingol et al., 2021). Although the strength of this connection varies by gene, we are able to rationalise biologically the relative success of certain pathways.

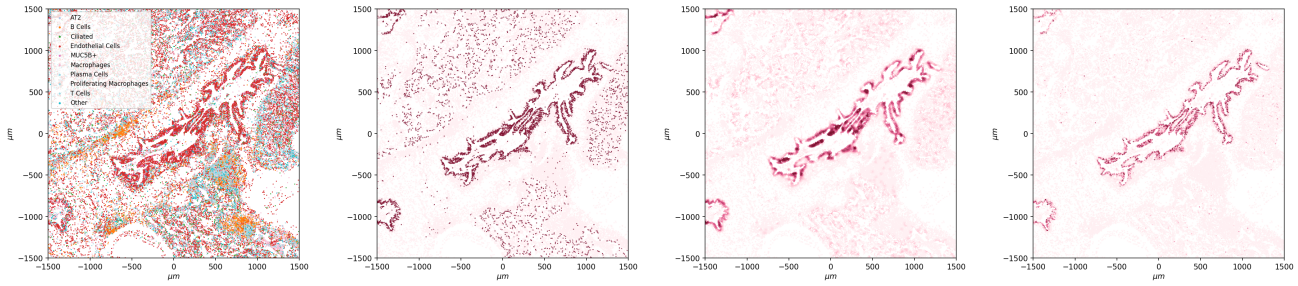
2. Related Work

The problem of unpaired translation has previously been explored for omics data. scConfluence (Samaran et al., 2024) used optimal transport to determine likely pairings across modalities, then trained encoders to align the latent embeddings based on these pairings. In contrast, scACT (Xu et al., 2024) trained encoders that map between modalities, namely scRNA-seq and ATAC-seq, enforced via a cycle consistency loss (Zhu et al., 2017). CellContrast (Li et al., 2024) designed a method for the recovery of a pseudo-space for single-cell data after contrastive learning on ST data. Spatial context is often modelled as an aggregation of expression values over the graph connecting neighbouring cells, for example by Hu et al. (2021); Birk et al. (2025); Wang et al. (2025). In contrast to the aforementioned methods, SCXM is trained to infer spatial context from single-cell data directly.

Generative pretraining of large transformer networks has acquired significant currency in omics analysis, following on from its successes in natural language processing and computer vision. Such models have come to be regarded as *foun-*

[1]GSK.ai, GlaxoSmithKline, London, United Kingdom. Correspondence to: Joseph Boyd <joseph.x.boyd@gsk.com>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.



(a) Cell types *(b)* Gene expression (binarised) *(c)* Spatially bulked *(d)* Model prediction

Figure 1. SCXM identifies single cells with high spatially bulked expression. Training is performed on dissociated single-cell data from different cell types (a), and performance is validated on high-plex spatial data, as depicted with: input gene expression for the *FOXJ1* gene (b), target spatially bulked expression (c), model predictions (d). It is seen that regions of dense gene expression and isolated expression in (b) are respectively promoted and attenuated in (c).

dational, given their aptitude for few-shot transfer to canonical single-cell applications, including cell typing, batch integration, and perturbation analysis (Yang et al., 2022; Cui et al., 2024; Hao et al., 2024). Geneformer (Theodoris et al., 2023) is a pioneering transformer for inference over transcriptomics data, combining a BERT (Devlin et al., 2019) transformer architecture with a non-parametric rank value encoder to order genes. The ordered `<gene_id>` embeddings combine with positional embeddings to impose a rank on genes, ordered by normalised expression. Nicheformer (Tejada-Lapuerta et al., 2025) extends Geneformer by joint training on spatial and non-spatial data.

3. Method

We denote scRNA-seq gene expression data $X \in \mathbb{R}^{|\mathcal{C}_{\text{SC}}|\times|\mathcal{G}_{\text{SC}}|}$ and ST expression data $Y \in \mathbb{R}^{|\mathcal{C}_{\text{ST}}|\times|\mathcal{G}_{\text{ST}}|}$, where $\mathcal{C}_{\text{SC}}$ is the set of cells sequenced under scRNA-seq with the gene panel $\mathcal{G}_{\text{SC}}$, and $\mathcal{C}_{\text{ST}}$ is the set of cells sequenced by ST with the gene panel $\mathcal{G}_{\text{ST}}$. In practice, $\mathcal{C}_{\text{SC}} \cap \mathcal{C}_{\text{ST}} = \emptyset$. That is, the cells of each modality originate in different tissue samples, or indeed patients. As such, the problem setting is one of unpaired translation. We only require that $\mathcal{G}_{\text{ST}} \cap \mathcal{G}_{\text{SC}} \neq \emptyset$, which in practice is always satisfied.

Our aim is to train a generative encoder model $G_{\text{E}} : X \to Z$, where Z is some representation of spatial context. Our approach will rely on two key assumptions: dissociated cells broadly preserve their formerly *in situ* gene expression when sequenced with a scRNA-seq modality; and ST and scRNA-seq count data, up to a straightforward normalisation, are interchangeable when measured across the same subset of genes. The first assumption will facilitate the learning of a cross-modal mapping, $G_{\text{E}} : X \to Z$, while the second assumption will be used to condition the encoder towards a consistent posterior, $G_{\text{E}}(X) \sim \text{P}_{Z|X}$.

3.1. Encoding Spatial Context

Cells in ST are typically modelled as nodes in a spatial graph (Palla et al., 2022), in which connectivity is determined on the basis of spatial proximity, either as a fixed number of nearest neighbours, or by a prescribed spatial radius. Given such a spatial graph, various measures exist for quantifying the concentration of gene expression in space, including spatial autocorrelation statistics (Moran, 1950; Geary, 1954). A natural approach to define a spatial context Z is via a graph convolution, which for a given cell i and gene g is computed as $z_{ig} = \sum_{j\in N(i)} k_{ij} y_{jg}$, where $N(i)$ is the neighbourhood of cell i, including a self-loop. The kernel weights k_{ij} may be apportioned in various ways, including uniform and Gaussian weights. In all experiments we use a Gaussian kernel with a radius of $25\mu m$.

3.2. Learning Framework

Our adopted learning framework is the Wasserstein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017). Wasserstein GANs (Arjovsky et al., 2017) use the earth mover's distance (EMD) as a metric for measuring the divergence between generated and ground truth distributions on data. The intractable EMD is reformulated using the Kantorovich-Rubinstein (KR) duality, yielding the training objective $\min_{G_{\text{E}}} \max_{D} \mathcal{L}_{\text{WGAN}}$, where

$$\mathcal{L}_{\text{WGAN}} = \mathbb{E}_{z\sim p_Z}\big[D(z)\big] - \mathbb{E}_{x\sim p_X}\big[D\big(G_{\text{E}}(x)\big)\big]. \quad (1)$$

A discriminator D is trained to estimate (in dual form) the EMD, which the generator G_{E} is trained to minimise. To satisfy the KR duality, the discriminator must be K-Lipschitz continuous, a condition that is enforced by a gradient penalty

$$\mathcal{L}_{\text{GP}} = \mathbb{E}_{\check{z}\sim P_{\check{Z}}}\big[\big(\|\nabla_{\check{z}} D(\check{z})\|_2 - 1\big)^2\big], \quad (2)$$

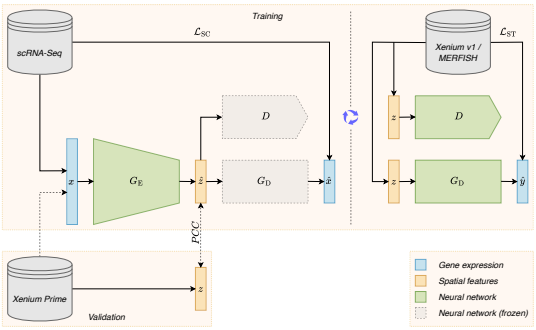


Figure 2. SCXM model architecture, with G_E a fine-tuned foundation model. The adversarial discriminator D enforces that inferred spatial features $\hat{z}$ remain in-distribution by approximating the earth mover's distance, while a decoder G_D enforces a consistent mapping. The decoder and discriminator are trained in alternation with the encoder. Training is performed on fully unpaired data sources, and high-plex spatial data serves as a paired validation set.

where $\tilde{z}$ are linearly interpolated samples of generated and ground truth data. We train a decoder G_D to reconstruct ST gene expression Y from precomputed Z deriving from the ST modality, subject to

$$\mathcal{L}_{ST} = \mathbb{E}_{y\sim p_Y}\left[\|G_D(z(y)) - y\|_2^2\right]. \quad (3)$$

Hence, the encoder G_E can be made consistent by decoding its predicted spatial features with the frozen decoder $\texttt{freeze}(G_D)$ by

$$\mathcal{L}_{SC} = \mathbb{E}_{x\sim p_X}\left[\|\texttt{freeze}(G_D)(G_E(x)) - x_{:\mathcal{G}_{ST}}\|_2^2\right], \quad (4)$$

where $x_{:\mathcal{G}_{ST}}$ denotes scRNA-seq expression sliced to the spatial panel of genes. In sum, our model resembles an adversarial autoencoder (Makhzani et al., 2015), trained in a semi-supervised way. The full objective is finally

$$\min_{G_E,G_D}\max_{D} \mathcal{L}_{WGAN} + \lambda_{GP}\cdot\mathcal{L}_{GP} + \lambda_{ST}\cdot\mathcal{L}_{ST} + \lambda_{SC}\cdot\mathcal{L}_{SC}, \quad (5)$$

where λ_{GP}, λ_{ST}, and λ_{SC} are tunable weights for the respective penalties. The model architecture is shown in Figure 2.

4. Experimental Setup

4.1. Datasets

Datasets were curated by combining public scRNA-seq and ST datasets according to tissue type. Our data sources were `CellxGene` (Program et al., 2025), the Gene Expression Omnibus (GEO) (Clough and Barrett, 2016), and 10x Genomics preview data (see Extended Table 2). Given the unpaired learning objective, a key challenge was validation. We approximated a paired validation set (see Figure 2) by using Xenium Prime data (10x Genomics, 2024), which is a ST protocol with a high-plex gene panel $\mathcal{G}_{PR}$ ($\sim$ 5000 genes). To ensure cross-modal compatibility, we slice the scRNA-seq and Xenium Prime data to $\mathcal{G}_{SC}\cap\mathcal{G}_{PR}$, and the ST data to $\mathcal{G}_{ST}\cap\mathcal{G}_{PR}$. We describe below the datasets analysed, and their dimensions after filtering and slicing.

Lung Dataset. The Human Lung Cell Atlas (HLCA) (Sikkema et al., 2023) scRNA-seq ($1,824,518$ cells $\times$ 4908 genes), Xenium V1 pulmonary fibrosis data (Vannan et al., 2025) ($1,093,782$ cells $\times$ 204 genes), and Xenium Prime for Human Lung Cancer (10x Genomics, 2025) ($159,771$ cells $\times$ 4908 genes).

Brain Dataset. Chromium of mouse brain nuclei (10x Genomics, 2022) ($44,725$ cells $\times$ 4980 genes), MERFISH mouse frontal cortex matter (Allen et al., 2023) ($367,582$ cells $\times$ 223 genes), and mouse hemispheric data with Xenium Prime (10x Genomics, 2024) ($56,799$ cells $\times$ 4980 genes). Mouse genes were mapped to their human homologues for compatibility with the vocabulary of the model backbone using `BioMart` (Smedley et al., 2009).

Breast Dataset. A scRNA-seq atlas of human breast cancer (Chen et al., 2026) ($404,579$ cells $\times$ 5054 genes), a Xenium V1 breast cancer cohort (Janesick et al., 2025) ($1,275,889$ cells $\times$ 148 genes), and Xenium Prime (10x Genomics, 2024) of a breast cancer ($107,004$ cells $\times$ 5054 genes).

4.2. Preprocessing

To foster consistency between the collected modalities, we performed cell filtering in a harmonised way across the spatial genes $\mathcal{G}_{ST}$, that is, the genes shared by all modalities. Cell type annotation was performed with CellTypist (Domínguez Conde et al., 2022) using tissue-appropriate models. Due to the significant domain shift remaining between the modalities, we adopted a simple binarisation binning technique for the reconstruction targets of the decoder G_D, that is, an indicator of non-zero gene expression. Graph convolutions were computed on the binarised expression with a customisation of `scanpy`'s (Wolf et al., 2018) metrics module.

4.3. Model Architecture

To bolster training on limited datasets, we utilised a pretrained GeneformerV2 (Chen et al., 2024) with 12 self-attention blocks as the backbone of the encoder G_E. To avoid inimical training dynamics, we froze all but the final four self-attention modules, and appended a two-layer MLP to the `<cls>` token to project to the target dimensionality.

A sigmoid activation function was used to ensure the projection reproduced the $[0, 1]$ range of the normalised spatial features. The default discriminator network D was a 5-layer MLP and G_{D} was a 4-layer MLP. As overfitting was observed on the significantly smaller brain dataset, we instead fully froze the backbone, and extended the projection head to three layers, while dropping a decoder layer. Throughout our experiments, we set $\lambda_{\mathrm{ST}} = \lambda_{\mathrm{SC}} = 1.0$ and $\lambda_{\mathrm{GP}} = 10.0$.

4.4. Baselines

As a means of comparison, we trained and evaluated two other omics translation models, scACT (Xu et al., 2024) and scConfluence (Samaran et al., 2024), on the task of translating between scRNA-seq gene expression and spatially bulked expression. As both methodologies were not originally trained on this novel task, we made the following adaptations to enable a reasonable comparison, and trained each on the same data splits as SCXM.

scACT. The assay for transposase-accessible chromatin using sequencing (ATAC-seq) branch was replaced with one that used spatially bulked expression. CellTypist (Domínguez Conde et al., 2022) annotations were used as the labels for the cell-type specific discriminator heads. Highly variable gene (HVG) selection was not applied to the scRNA-seq input to maintain parity with SCXM.

scConfluence. The inter-modality cost matrix required by the inverse optimal transport loss was computed via correlation distance on the log-normalised spatial panel $\mathcal{G}_{\mathrm{ST}}$ in both the scRNA-seq and ST data. Binary cross-entropy was used as the reconstruction loss for the spatially bulked expression. HVG selection was not applied.

5. Results

5.1. Cross-modal Translation

SCXM outperformed prior methods for unpaired translation across all three curated datasets, as shown in Figure 3, with median Pearson correlation coefficient (PCC) of 0.4291, 0.4551, and 0.1700 on the lung, breast, and brain datasets, respectively, compared with 0.1750, 0.1964, and 0.0573 for scConfluence, and 0.0065, 0.0284, and 0.0151 for scACT. We observed inferior performances on the significantly smaller brain dataset, although the best performing gene for SCXM scored a PCC of 0.7085, comparable with 0.7050 and 0.6686 in the lung and breast datasets. We postulate that the superior performance of SCXM for ST to scRNA-seq transfer may be attributed to its use of consistency penalties exploiting the count data relation between the modalities (Equations 3 and 4), as this amounts to semi-supervision, in contrast with the unsupervised cycle consistency of scACT and optimal transport plan of scConfluence.

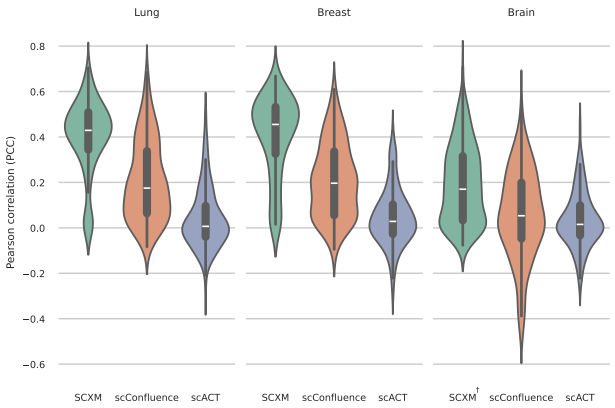


Figure 3. Gene-wise correlations on Xenium Prime validation data for cross-modal translation methods across three datasets. † On the smaller brain dataset, SCXM was trained with a frozen GeneformerV2 with MLP head.

5.2. Niche Characterisation

Predicting context as spatially bulked gene expression allows us to separate clustered expression from isolated expression, as shown in Figure 1 and Extended Figures 5-8. Such readouts may provide hints of the presence of niches in the tissue. We observe in Extended Figure 6 an enrichment of B cells in regions predicted by SCXM to contain a high spatially bulked expression of *CXCR4*. By thresholding the prediction at 0.5, we can segment B cells across the lung tissue into isolated and clustered groups (Figure 4). Applying differential analysis, we observe a >4 log-fold change for *POU2AF1*, *MS4A1*, and *CD79A* ($p_{\mathrm{adj}} < .001$, Benjamini-Hochberg). *POU2AF1* in particular is implicated in the formation of biological niches such as germinal centres (Tiedt et al., 2001).

5.3. Gene Set Enrichment Analysis

In the lung dataset, we investigated whether SCXM predictions recapitulate certain aspects of spatial biology better than others. We selected a set of 25 pathways relevant to lung tissue and subset to those genes which were measured within the spatial panel. A gene set enrichment analysis (GSEA) was performed over these pathways using PCC as the ordinal metric, with a normalised enrichment score and Benjamini-Hochberg FDR correction used as in (Subramanian et al., 2005). Results are shown in Table 1. The top two pathways, related to lung lineage and angiogenesis, were the most significant, highlighting the ability of SCXM to capture larger scale spatial biology. Six more borderline-significant pathways dealt largely with immune response, representing a more localised aspect of spatial biology at the epithelium-immune interface. Pathways related to cell-cell signalling and cell cycle were more difficult to recapitulate, perhaps due to fewer spatial constraints governing these pro-

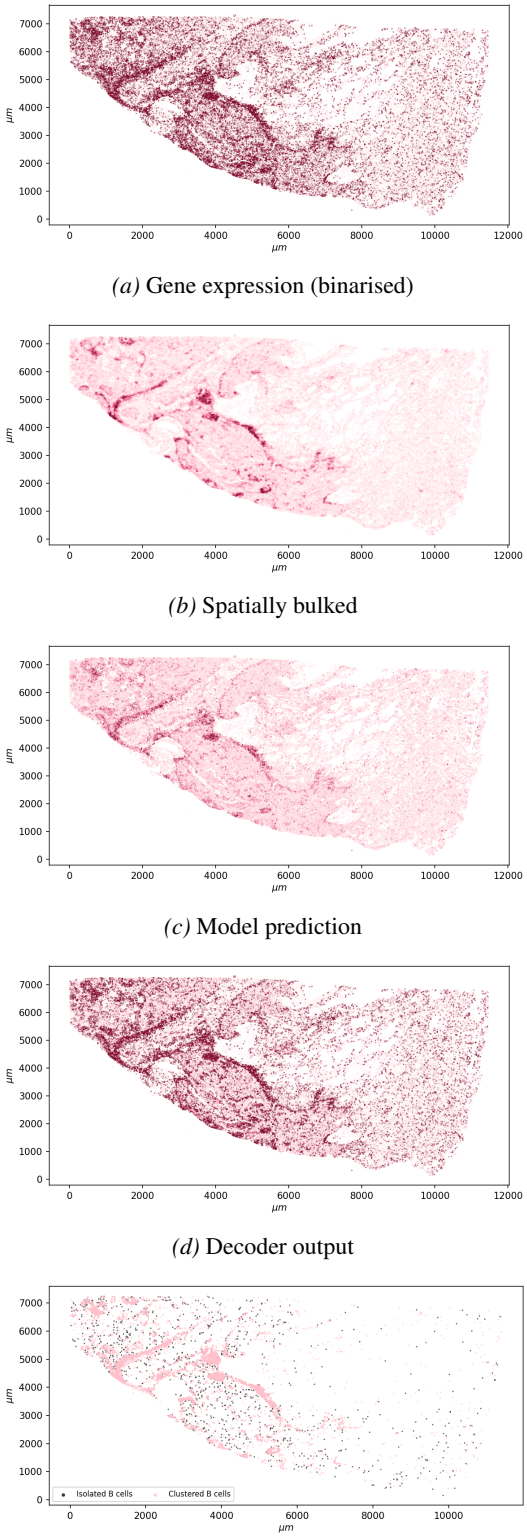
cesses. Thus, initial results indicate that the main strength of SCXM may intuitively lie in its ability to capture more prominent spatial context, and future work is planned to better understand its limits in this regard.

Table 1. GSEA results for the lung dataset. NES: normalised enrichment score; FDR: false discovery rate (Benjamini-Hochberg).

Pathway	Size/Overlap	NES	p	FDR
Lung lineage identity	50/15	1.835	<0.001	0.003
Angiogenesis	36/18	1.894	<0.001	0.003
T cell activation	100/26	1.328	0.010	0.063
Myeloid innate immunity	130/17	1.106	0.009	0.063
Interferon response	200/9	1.419	0.013	0.067
Immune checkpoint	80/12	1.373	0.025	0.099
Hedgehog signalling	47/7	1.431	0.028	0.099
Notch signalling	48/11	1.450	0.055	0.171
Integrin cell adhesion	199/13	1.276	0.074	0.206
B cell markers	71/13	1.173	0.092	0.229
Unfolded protein response	113/14	0.987	0.153	0.348
Chemokine signalling	190/10	1.136	0.185	0.386
PI3K/AKT/mTOR survival	354/15	0.976	0.316	0.607
EMT	200/21	1.224	0.340	0.607
Hypoxia response	200/16	0.958	0.457	0.761
EGFR signalling	85/12	0.977	0.563	0.880
FGF signalling	55/11	1.050	0.656	0.965
Apoptosis	136/14	1.010	0.906	1.000
NF-κB signalling	100/13	0.977	0.986	1.000
Cell cycle / proliferation	124/13	0.988	0.762	1.000
Hippo / YAP-TAZ	163/13	0.970	0.957	1.000
KEGG NSCLC pathway	66/17	1.070	0.912	1.000
Oxidative stress / NRF2	100/11	0.976	0.780	1.000
TGF-β signalling	94/17	−0.856	0.952	1.000
WNT signalling	158/19	−0.573	1.000	1.000

6. Conclusion

With SCXM we have demonstrated preliminary success at cross-modal translation of scRNA-seq and ST data from completely unpaired data sources. Performance varied with dataset size, and future extensions could explore training on larger datasets. Moreover, assembling suitable dataset combinations spanning three modalities for the training and validation of SCXM proved to be challenging. Even when correspondence exists between the high-level properties of the studied cohorts, there is non-negligible domain shift across the modalities that makes model optimisation and validation difficult. One solution could be to move to a pan-tissue, pan-modal dataset, such as in (Tejada-Lapuerta et al., 2025), which could allow for model generalisation beyond study-specific biases. A challenge for SCXM would then be to integrate the disjoint spatial gene panels, which can vary significantly by technology and tissue type. We thus posit that future spatial niche modelling with SCXM could incorporate spatial cell type composition in addition to, or instead of, spatially bulked gene expression, facilitating a more universal representation of spatial biology.



(a) Gene expression (binarised)

(b) Spatially bulked

(c) Model prediction

(d) Decoder output

(e) Spatial subtyping

Figure 4. Whole-slide view of B cell spatial subtyping from *CXCR4* gene expression input (a), target spatially bulked readout (b), model encoder prediction (c), model decoder output (d), and thresholded model outputs (e) as decision rule for distinguishing isolated B cells (black) and clustered B cells (pink).

Acknowledgements

We extend our thanks to Lefteris Zormpas, Stefan Groha, and Richard Li for providing helpful feedback and discussions.

Impact Statement

The objective of this work is to better understand the connections between different single-cell modalities, in an effort to better understand fundamental biology and characteristics of disease revealed by gene expression. This work could therefore have a general scientific impact as well as an impact on the development of future therapies.

References

10x Genomics. Mouse brain nuclei isolated with chromium nuclei isolation kit, saltyez protocol, and 10x complex tissue dp (ct sorted and ct unsorted). Technical Report CG000552, 10x Genomics, 2022.

10x Genomics. Xenium in situ gene expression - post-xenium analyzer h&e staining. Technical Report CG000613, 10x Genomics, 2024.

10x Genomics. Post-xenium in situ applications: Immunofluorescence, h&e, visium v2, and visium hd. Technical Report CG000709, 10x Genomics, 2025.

W. E. Allen, T. R. Blosser, Z. A. Sullivan, C. Dulac, and X. Zhuang. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1):194–208, 2023.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. Pmlr, 2017.

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

S. Birk, I. Bonafonte-Pardàs, A. M. Feriz, A. Boxall, E. Agirre, F. Memi, A. Maguza, A. Yadav, E. Armingol, R. Fan, et al. Quantitative characterization of cell niches in spatially resolved omics data. *Nature Genetics*, 57(4):897–909, 2025.

A. Chen, L. Kroehling, C. S. Ennis, G. V. Denis, and S. Monti. A highly resolved integrated single-cell atlas of human breast cancers. *NAR Genomics and Bioinformatics*, 8(1):lqaf217, 2026.

H. Chen, M. S. Venkatesh, J. G. Ortega, S. V. Mahesh, T. N. Nandi, R. K. Madduri, K. Pelka, and C. V. Theodoris. Quantized multi-task learning for context-specific representations of gene network dynamics. *bioRxiv*, 2024.

K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

E. Clough and T. Barrett. The gene expression omnibus database. In *Statistical genomics: methods and protocols*, pages 93–110. Springer, 2016.

H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. Howlett, O. Suchanek, K. Polanski, H. King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

R. C. Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11): 1342–1351, 2021.

A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1):8353, 2023.

A. Janesick, S. N. Kravitz, W. Stauffer, M. Valencia, and S. E. Taylor. Biomarker quantification in breast cancer using xenium in situ. *bioRxiv*, pages 2025–12, 2025.

S. Li, J. Ma, T. Zhao, Y. Jia, B. Liu, R. Luo, and Y. Huang. Cellcontrast: Reconstructing spatial relationships in single-cell rna sequencing data via deep contrastive learning. *Patterns*, 5(8), 2024.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

G. Palla, H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

C. C. S. Program, S. Abdulla, B. Aevermann, P. Assis, S. Badajoz, S. M. Bell, E. Bezzi, B. Cakir, J. Chaffer, S. Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic acids research*, 53 (D1):D886–D900, 2025.

J. Samaran, G. Peyré, and L. Cantini. scconfluence: single-cell diagonal integration with regularized inverse optimal transport on weakly connected features. *Nature Communications*, 15(1):7762, 2024.

L. Sikkema, C. Ramírez-Suástegui, D. C. Strobl, T. E. Gillett, L. Zappia, E. Madissoon, N. S. Markov, L.-E. Zaragosi, Y. Ji, M. Ansari, et al. An integrated cell atlas of the lung in health and disease. *Nature medicine*, 29(6): 1563–1577, 2023.

D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk. Biomart–biological queries made easy. *BMC genomics*, 10(1):22, 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

A. Tejada-Lapuerta, A. C. Schaar, R. Gutgesell, G. Palla, L. Halle, M. Minaeva, L. Vornholz, L. Dony, F. Drummer, T. Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nature methods*, pages 1–14, 2025.

C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

R. Tiedt, B. A. Bartholdy, G. Matthias, J. W. Newell, and P. Matthias. The ring finger protein siah-1 regulates the level of the transcriptional coactivator obf-1. *The EMBO journal*, 20(15):4143–4152, 2001.

A. Vannan, R. Lyu, A. L. Williams, N. M. Negretti, E. D. Mee, J. Hirsh, S. Hirsh, N. Hadad, D. S. Nichols, C. L. Calvi, et al. Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nature genetics*, 57(3):647–658, 2025.

C. Wang, H. Cui, A. Zhang, R. Xie, H. Goodarzi, and B. Wang. scgpt-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. *biorxiv*, pages 2025–02, 2025.

F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

S. Xu, J. Liu, and J. Zhang. scact: accurate cross-modality translation via cycle-consistent training from unpaired single-cell data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2722–2731, 2024.

F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature machine intelligence*, 4(10):852–866, 2022.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

A. Extended Tables

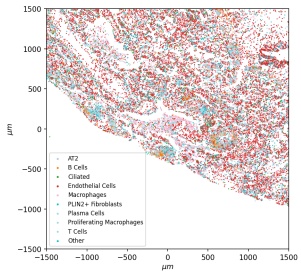
A summary of the datasets used for training and validating SCXM is provided in Table 2.

B. Extended Figures

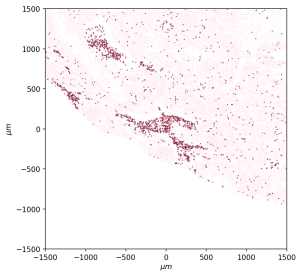
Figures 5, 6, 7, and 8 provide illustrative examples of SCXM inputs, targets, and predictions. Regions of dense prediction often correspond to clusters enriched in a given cell type, for example macrophages (Figure 5) and B cells (Figure 6).

Table 2. Summary of datasets used for training and validating SCXM.

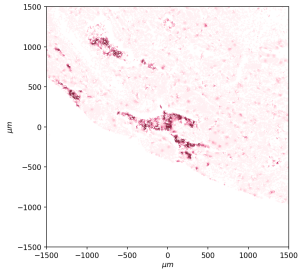
Dataset	Modality	Title	Retrieved from
Lung (human)	scRNA-Seq	An integrated cell atlas of the lung in health and disease (Sikkema et al., 2023)	https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293
	Xenium v1	Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis (Vannan et al., 2025)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276945
	Xenium Prime	Post-Xenium Technical Note: Xenium v1 and Xenium Prime 5K for FFPE Human Lung Cancer (10x Genomics, 2025)	https://www.10xgenomics.com/datasets/xenium-human-lung-cancer-post-xenium-technote
Brain (mouse)	scRNA-Seq	Mouse Brain Nuclei Isolated with Chromium Nuclei Isolation Kit (10x Genomics, 2022)	https://www.10xgenomics.com/datasets/mouse-brain-nuclei-isolated-with-chromium-nuclei-isolation-kit-saltyez-protocol-and-10x-complex-tissue-dp-ct-sorted-and-ct-unsorted-1-standard
	MERFISH	Molecular and spatial signatures of mouse brain aging at single-cell resolution (Allen et al., 2023)	https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e
	Xenium Prime	Fresh Frozen Mouse Brain Hemisphere with 5K Mouse Pan Tissue and Pathways Panel (10x Genomics, 2024)	https://www.10xgenomics.com/datasets/xenium-prime-fresh-frozen-mouse-brain
Breast (human)	scRNA-Seq	A highly resolved integrated single-cell atlas of human breast cancers (Chen et al., 2026)	https://cellxgene.cziscience.com/collections/9432ae97-4803-4b9f-8f64-2b41e42ad3cb
	Xenium v1	Biomarker Quantification in Breast Cancer using Xenium In Situ (Janesick et al., 2025)	https://www.10xgenomics.com/datasets/xenium-ffpe-human-breast-biomarkers
	Xenium Prime	FFPE Human Breast Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 Custom Genes (10x Genomics, 2024)	https://www.10xgenomics.com/datasets/xenium-prime-ffpe-human-breast-cancer



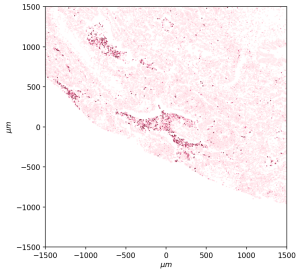
(a) Cell types



(b) Gene expression (binarised)

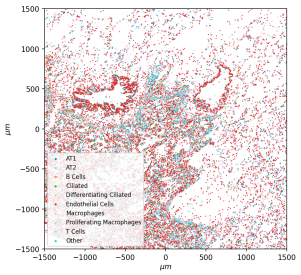


(c) Spatially bulked

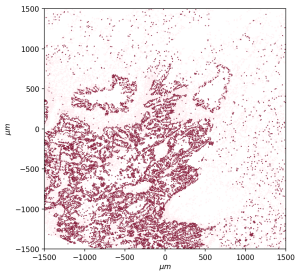


(d) Model prediction

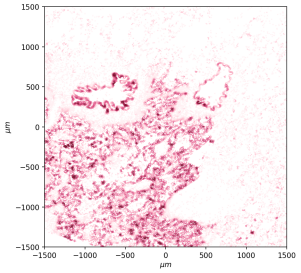
Figure 5. Sample region with cell types (a), gene expression inputs (b), targets (c) and model prediction (d) for *MARCO* gene.



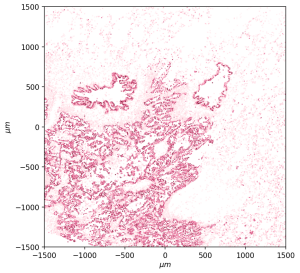
(a) Cell types



(b) Gene expression (binarised)

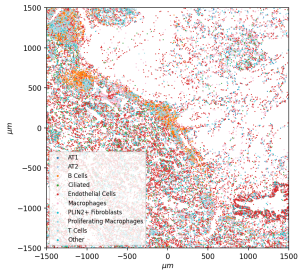


(c) Spatially bulked

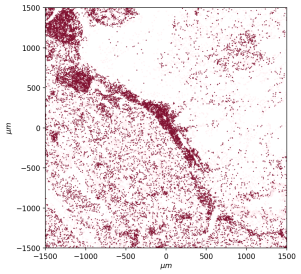


(d) Model prediction

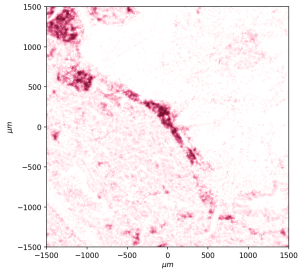
Figure 7. Sample region with cell types (a), gene expression inputs (b), targets (c) and model prediction (d) for *EPCAM* gene.



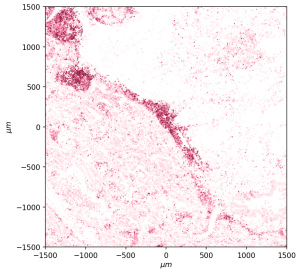
(a) Cell types



(b) Gene expression (binarised)

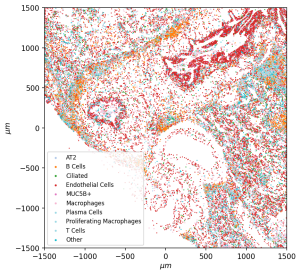


(c) Spatially bulked

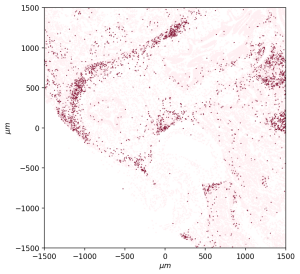


(d) Model prediction

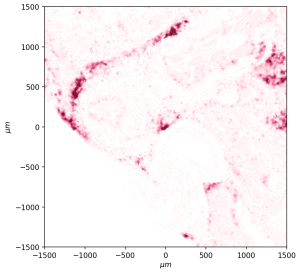
Figure 6. Sample region with cell types (a), gene expression inputs (b), targets (c) and model prediction (d) for *CXCR4* gene.



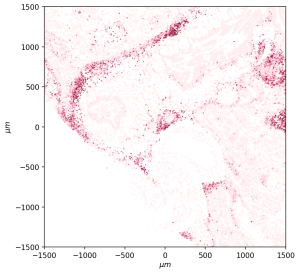
(a) Cell types



(b) Gene expression (binarised)



(c) Spatially bulked



(d) Model prediction

Figure 8. Sample region with cell types (a), gene expression inputs (b), targets (c) and model prediction (d) for *TNFRSF13C* gene.