Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MuSpec: an unsupervised multi-scale spatial-niche encoder recovers conserved granuloma niches and nominates host-directed repurposing candidates in human pulmonary tuberculosis

Yuxuan Gou
Xinjiang Second Medical College
Tingting Luo
Xinjiang Second Medical College
Chunhong Xiao
Xinjiang Second Medical College
Jie Chen
Xinjiang Second Medical College
Yating Zhao
Xinjiang Second Medical College
Yang Yang
Xinjiang Second Medical College
Jing Chen
Xinjiang Second Medical College
Yi Peng
The Sixth People's Hospital of Xinjiang Uygur Autonomous Region, The Infectious Disease Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region, China
Jiangdong Wu
1556874645@qq.com

Xinjiang Second Medical College https://orcid.org/0000-0002-4804-9706
Qifeng Li
Xinjiang Hospital of Beijing Children's Hospital, Children's Hospital of Xinjiang Uygur Autonomous Region, The Seventh People's Hospital of Xinjiang Uygur Autonomous Region CN

Research Article

Keywords: spatial transcriptomics, Stereo-seq, tuberculosis granuloma, spatial niche, multi-scale deep learning, external attention, host-directed therapy, drug repurposing, molecular docking, human lung

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11339003/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

MuSpec: an unsupervised multi-scale spatial-niche encoder recovers conserved granuloma niches and nominates host-directed repurposing candidates in human pulmonary tuberculosis

Draft original-research manuscript — target journal: *Journal of Translational Medicine* (BMC).

Authors: Yuxuan Gou[1], Tingting Luo[1], Chunhong Xiao[1], Jie Chen[1], Yating Zhao[1], Yang Yang[1], Jing Chen[1], Yi Peng[3]„ *Jiangdong Wu*[1],, Qifeng Li[2],*

Affiliations: [1] Xinjiang Second Medical College, Xinjiang, China. [2] Xinjiang Hospital of Beijing Children's Hospital / Children's Hospital of Xinjiang Uygur Autonomous Region / Seventh People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China. [3] Sixth People's Hospital of Xinjiang Uygur Autonomous Region / Infectious Disease Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China.

Yuxuan Gou, Tingting Luo and Chunhong Xiao contributed equally to this work. Yi Peng, Jiangdong Wu and Qifeng Li are corresponding authors.

Corresponding authors: Yi Peng (pengyi4198@163.com); Jiangdong Wu (1556874645@qq.com); Qifeng Li (liqiqi521@sina.com).

Manuscript status note. All quantitative values in this draft are transcribed from the project result tables (Tables 1–7, Supplementary Tables S1–S3) and from the six main figures.

Abstract

Background. Tuberculosis (TB) granulomas are spatially organised host–pathogen structures, and the spatial arrangement of immune and stromal cells within the granuloma shapes disease outcome. Spatial transcriptomics can now resolve these structures at single-cell resolution, but most spatial-niche models have been developed and validated on tumour tissue. There is a shortage of unsupervised methods benchmarked on infectious-disease granulomas, and little evidence on how such niche models behave under input noise, missing modalities, and cross-sample transfer — the conditions that decide whether a niche atlas is translational or merely descriptive.

Methods. We assembled **MuSpec**, an unsupervised multi-scale spatial-niche encoder, and applied it to two single-cell-resolution Stereo-seq slides from human TB lung: **B1** (pulmonary granuloma; Y40012NC13; 36,382 cells) and **P1** (peri-lesional lung; Y40012LA11; 51,834 cells). Each cell was represented as a 39-dimensional input (7-dimensional cell-type composition from cell2location deconvolution against a 146,345-cell peripheral-blood scRNA-seq reference, 30 expression principal components, and 2 spatial coordinates). MuSpec smooths the input at three kNN scales (k = 10/20/30), rasterises each scale onto a $\leq 224^2$ lattice, and passes the three feature maps through four deep-learning modules built on published attention-and-fusion designs — ASFF cross-scale fusion, TripletAttention, ExternalAttention, and a Fourier graph network (FGN) — to produce a 16-dimensional niche embedding; a shared encoder is trained jointly across both slides with an MSE reconstruction loss plus a spatial-smoothing loss, and niches are obtained by KMeans (K = 9). We benchmarked MuSpec against BANKSY at matched numbers of spatial domains, ran a three-seed module ablation and a 34-configuration single-factor hyperparameter sweep, and used the recovered granuloma signature for multi-library signature-reversal drug repurposing (LINCS L1000 chemical perturbations, GEO perturbations, NIBR DRUGseq 2025, and the original CMAP, merged) and AutoDock Vina docking against the host target SP110.

Results. MuSpec produced a compact, sub-million-parameter model (782,251 parameters; 93.9% in ASFF) that trains in minutes on a single consumer GPU (≈7 GB peak). On every matched-K comparison, MuSpec embeddings separated niches far more sharply than BANKSY (Δsilhouette +0.31 to +0.40 on both slides); MuSpec's own K = 9 silhouette was 0.322 (B1) and 0.285 (P1). Three-seed ablation showed that **no module was removable losslessly**: removing ExternalAttention lowered niche separation (Δsilhouette −0.093) but raised robustness and cross-slide transfer (Δ of +0.137 to +0.362), identifying a clean separation↔robustness trade-off. For normalisation, a three-seed paired design showed that 7 of 8 robustness metrics changed sign across seeds, so we report **no seed-stable difference** between global and per-slide normalisation, with only a small stable (+0.010 ± 0.003) silhouette advantage to per-slide normalisation on B1. The hyperparameter sweep revealed a monotone separation–robustness frontier with no cliffs, on which the default setting was the robustness optimum (mean noise-robustness ARI 0.758; drop-composition ARI 0.683 on B1 and 0.518 on P1, the highest in the sweep). Joint niche characterisation recovered myeloid/granuloma (SPP1, MMP9, CTSK), airway/club (MUC5B, SCGB1A1, BPIFB1), fibrotic (SFRP2, COL1A1, INHBA) and fibroblast (IL11, TFPI2, PI16) niches, and an SP110–IFN hub analysis. A 324-gene granuloma signature (233 up, 91 down) yielded top

signature-reversal hits led by a **PPARγ-agonist trio** (pioglitazone 90.08, troglitazone, rosiglitazone), and in-silico docking of a focused 6-ligand panel against SP110 gave the strongest affinities for withaferin A (−7.18 kcal/mol) and nintedanib (−6.92 kcal/mol).

Conclusions. MuSpec provides an unsupervised, reproducible, computationally light route to spatially resolved niche atlases in infectious-disease tissue, and its systematic benchmarking exposes a separation↔robustness trade-off that should be reported rather than hidden. Applied to human TB lung, it recovered biologically coherent, cross-slide-conserved granuloma niches and nominated host-directed repurposing candidates.

Keywords: spatial transcriptomics; Stereo-seq; tuberculosis granuloma; spatial niche; multi-scale deep learning; external attention; host-directed therapy; drug repurposing; molecular docking; human lung.

Background

Tuberculosis (TB) remains a leading cause of death from a single infectious agent, and the outcome of infection is decided largely within the **granuloma**, the organised cellular structure that the host builds around *Mycobacterium tuberculosis* (2). The granuloma is not a homogeneous ball of immune cells: macrophage subsets, T-cell states, B cells, and fibroblast-like stromal cells are arranged in reproducible spatial zones, and the signalling that sustains or fails to contain infection is itself spatially organised (4, 5). This spatial architecture matters clinically, because host-directed therapies that aim to modulate this architecture rather than to kill the bacterium directly depend on knowing which cell states sit in which microenvironment (6). A descriptive map of the granuloma is therefore not a convenience; it is the substrate on which host-directed treatment hypotheses are built.

Spatial transcriptomics has made such maps technically feasible. Single-cell-resolution platforms, including Stereo-seq DNA-nanoball arrays, resolve both transcript identity and tissue coordinates at subcellular scale across whole tissue sections (7). The bottleneck has shifted from measurement to modelling: given tens of thousands of spatially resolved cells, we need to define the recurrent **spatial niches** — self-consistent local microenvironments — in a way that is reproducible, comparable across samples, and robust to the imperfections of real data.

A large family of computational methods now addresses spatial-domain detection. SpaGCN couples gene expression, spatial location, and histology through a graph convolutional network (8); STAGATE uses an adaptive graph attention auto-encoder to identify spatial domains (9); BANKSY unifies cell-type clustering and tissue-domain segmentation by augmenting a cell's own expression with its neighbourhood (10). More recent work has moved toward learned niche representations and foundation models: NicheCompass learns interpretable niche embeddings (11), Novae provides a scalable foundation model for spatial-domain segmentation (12), and GITIII addresses cross-sample and cross-modal spatial integration (13); scPerturBench, in the perturbation-modelling setting, sets out benchmarks that enforce fair, matched comparison rather than favourable defaults (15). These methods have been developed and demonstrated predominantly on tumour tissue and on spot-based or lower-resolution assays. Infectious-disease tissue at single-cell resolution is a different regime: granulomas are small, intensely heterogeneous, and are usually available in only a handful of slides, so methods must perform with very few samples and must be explicitly stress-tested under the noise and missing-modality conditions that these datasets routinely present.

The closest prior art to the present work is **SpaCEy** (14), which builds a spatial-niche model for tumour tissue. SpaCEy demonstrates that learned niche representations are a productive abstraction, but it is aimed at tumour ecosystems and, as we read it, relies on a supervision scheme that assumes annotated or otherwise informed niche structure. What is missing for infectious disease is an **unsupervised** niche encoder that (i) is validated on granuloma, not tumour, tissue; (ii) is benchmarked at matched numbers of domains against a strong geometry-based baseline rather than against favourable defaults; (iii) reports its behaviour under input noise, missing modalities, and cross-slide transfer with seed-controlled statistics; and (iv) is light enough that small academic groups without multi-GPU infrastructure can reproduce the whole analysis.

Here we present **MuSpec**, an unsupervised multi-scale spatial-niche encoder, and apply it to two single-cell-resolution Stereo-seq slides from human TB lung — a pulmonary granuloma (B1) and peri-lesional lung tissue from the same source (P1). MuSpec is deliberately modest in design: it composes four functional modules whose architectures follow published attention-and-fusion designs within a multi-scale pipeline that we implement ourselves (data loading, multi-scale graph construction, rasterisation, projection, and the wiring that joins the modules). The contribution of the paper is fivefold, and we state it in detail because the novelty is not a single module but the combination. **First, a multi-scale, two-modality cell representation.** Each cell is described not by its own state alone but by its state at three spatial neighbourhood scales (k = 10/20/30,

spanning roughly tens to hundreds of neighbours) and by two orthogonal modalities — a deconvolution-derived cell-type composition block and an expression-PC block — bound to spatial coordinates. This makes the per-cell token a local-microenvironment descriptor rather than a single-cell profile, which is the representation a niche model requires; a single-scale variant on the same input loses separation (Δsilhouette -0.040 ± 0.031, seed-stable), so the multi-scale construction is load-bearing rather than decorative. **Second, an explicit multi-scale fusion-and-attention wiring.** The three scales are rasterised onto a shared lattice, projected, and adaptively fused (ASFF) with learned, location-specific weights before a channel-memory external attention and a Fourier graph encoder compress them into a 16-dimensional embedding; that embedding is read back into the cells through a differentiable grid_sample at their fractional coordinates, so the pipeline is end-to-end differentiable and order-invariant. We bound this claim honestly: the four functional modules implement published attention-and-fusion architectures, and our methodological contribution is the multi-scale wiring, the differentiable readout, and the joint-training loop. **Third, joint cross-slide training with an explicit spatial-smoothness prior.** A single shared encoder is trained on both slides — a granuloma and its peri-lesional margin — with a loss that adds a kNN-smoothing term on the embedding itself, so niche identity is shared across slides by construction and the embedding is constrained to be a smooth field over the tissue rather than an unconstrained per-cell latent; the leave-one-dataset-out protocol then tests that sharing directly rather than assuming it. **Fourth, a benchmark and stress-test protocol that reports a frontier rather than a winner.** We benchmark at **matched numbers of domains** against BANKSY, and we run a three-seed module ablation and a 34-configuration single-factor sweep; these independently expose an empirical **separation↔robustness frontier** on which the default configuration sits at the robust end, and they include a seed-stable result that external-memory attention trades separation for invariance. This is the part of the work we consider most useful and most transferable, because it is the part a single-seed, single-configuration study would have suppressed. **Fifth, a translational read-out for infectious-disease tissue.** We use the recovered granuloma signature for LINCS L1000 signature reversal, nominating a mechanistically coherent PPARγ-agonist set, and add in-silico docking against the host target SP110 supported by orthogonal cross-checks. Together these five contributions answer a specific gap — a light, unsupervised, honestly benchmarked multi-scale niche encoder developed and validated on granuloma rather than tumour tissue — and the remainder of the paper substantiates each in turn.

Methods

Data and preprocessing

Two human lung Stereo-seq cellbin slides were analysed (Table 1): **B1** (Y40012NC13) from a pulmonary granuloma (36,382 cells) and **P1** (Y40012LA11) from peri-lesional lung tissue of the same source (51,834 cells). Cell counts are the intersection of deconvolution-scored barcodes, spatial coordinates, and expression, and differ by ±2 from the raw cellbin totals (36,384 and 51,835). For each slide, raw counts were normalised to counts-per-10,000 (CP10K), log1p-transformed, restricted to 2,000 highly variable genes (HVG), and reduced to 30 principal components by randomised SVD with seed 0 (the expression-PC block). Spatial coordinates supplied the 2-dimensional spatial block. The composition block (7 dimensions) was obtained by cell2location deconvolution against a 146,345-cell peripheral-blood scRNA-seq reference spanning 5 healthy-control, 5 drug-sensitive (YM = DS), and 5 drug-resistant (NY = DR) donors, using the seven shared cell-type classes (T, dendritic, NK, $CD8^+$ T, B, $CD4^+$ T, monocyte). The per-cell input to MuSpec was the concatenation $7 \Vert 30 \Vert 2 = 39$ dimensions.

MuSpec architecture

MuSpec maps each cell to a 16-dimensional niche embedding through the pipeline in Table 2a and Fig. 1. The 39-dimensional per-cell input is smoothed by kNN averaging at three neighbourhood scales (k = 10, 20, 30), producing three feature sets. Each is rasterised onto a $\leq 224^2$ lattice by bin-averaging, yielding three image tensors of shape (1, 39, H, W) — batch size 1, 39 channels (the modality blocks are stacked as channels, not as a single channel). Each raster is projected by a 1×1 convolution (to 512, 256, and 256 channels for the three scales). Average-pooling (×4 on the coarsest scale, ×2 on the next) aligns the three scales, and ASFF(level = 2) fuses them into a 256-channel map. TripletAttention(256) applies spatial attention across the (H, W, C) dimensions. A 1×1 convolution reduces the map to 32 channels, and grid_sample samples each cell's 32-dimensional feature at its fractional grid coordinate. ExternalAttention(32, 8 heads) applies channel-wise external-memory attention, and FGN(pre_length=16, embed_size=8, feature_size=1, seq_length=32, hidden_size=32) produces the 16-dimensional niche embedding. A linear layer decodes the embedding back to 39 dimensions. The architecture additionally carries an optional linear read-out head ($16 \rightarrow 1$) from the module library that maps the embedding to a per-cell logit; this head is retained for completeness and for potential supervised downstream tasks, but it is **not** part of the unsupervised objective used here — the reported MuSpec is trained with the reconstruction and spatial-smoothing losses only, and no per-cell labels are used at any stage.

The four functional modules (ASFF, TripletAttention, ExternalAttention, FGN) implement published attention-and-fusion architectures, and all code in the pipeline (data loading, composition block, multi-scale kNN smoothing, rasterisation, projections, pyramid, fuse head, decoder, joint-training loop, normaliser, augmentation, and robustness harness) was written for this study (Table 2b). An import-time compatibility shim for an optional accelerated-attention package and a CPU fallback for CUDA calls are environment adapters only and do not alter module behaviour on GPU.

Mathematical formulation

The forward pass, loss, and downstream statistics are stated here as explicit formulas; symbols follow Table 2a. On slide s with N_s cells, cell i carries the 39-dimensional input

$$\boldsymbol{x}_i = [\,\boldsymbol{c}_i \;\|\; \boldsymbol{p}_i \;\|\; \boldsymbol{u}_i\,] \in \mathbb{R}^{39}, \qquad \boldsymbol{c}_i \in \mathbb{R}^{7}, \quad \boldsymbol{p}_i \in \mathbb{R}^{30}, \quad \boldsymbol{u}_i = (x_i, y_i) \in \mathbb{R}^{2},$$

where $\boldsymbol{c}_i$ is the cell2location composition block, $\boldsymbol{p}_i$ the expression-PC block, and $\boldsymbol{u}_i$ the spatial coordinate.

Multi-scale kNN smoothing. For each scale $k \in \{10,20,30\}$ let $N_k(i)$ be the k spatial nearest neighbours of cell i (Euclidean distance in $\boldsymbol{u}$, self excluded). The smoothed input is the neighbour mean

$$\tilde{\boldsymbol{x}}_i^{(k)} = \frac{1}{k} \sum_{j \in N_k(i)} \boldsymbol{x}_j.$$

Rasterisation by bin-averaging. With bin size b, cell i falls in grid bin (r_i, c_i) given by

$$r_i = \mathrm{clip}(\lfloor y_i - y_{\min}/b \rfloor, 0, H-1), \qquad c_i = \mathrm{clip}(\lfloor x_i - x_{\min}/b \rfloor, 0, W-1).$$

Each scale yields one image tensor $\boldsymbol{G}^{(k)} \in \mathbb{R}^{1\times 39\times H\times W}$ whose bin value is the average of the smoothed vectors landing in that bin:

$$\boldsymbol{G}^{(k)}[0,:,r,c] = \frac{1}{\max(n_{rc},1)} \sum_{i:\,(r_i,c_i)=(r,c)} \tilde{\boldsymbol{x}}_i^{(k)}, \qquad n_{rc} = \#\{\, i : (r_i,c_i) = (r,c) \,\}.$$

Scale projection and pooling pyramid. Three 1×1 convolutions project the rasters to C_k channels, and average-pooling aligns them to a common lattice:

$$\boldsymbol{F}_k = \mathrm{Conv}_{1\times 1}(\boldsymbol{G}^{(k)};\, C_k), \qquad (C_{30}, C_{20}, C_{10}) = (512,256,256),$$

$$\boldsymbol{L}_0 = \mathrm{AvgPool}_{4\times4}(\boldsymbol{F}_{30}), \qquad \boldsymbol{L}_1 = \mathrm{AvgPool}_{2\times2}(\boldsymbol{F}_{20}), \qquad \boldsymbol{L}_2 = \boldsymbol{F}_{10}.$$

Adaptive cross-scale fusion (ASFF). A small convolutional sub-network produces a per-scale, per-location weight map $\alpha_k \in (0,1)^{H\times W}$, normalised across the three levels; the levels are up-sampled by bilinear interpolation to the finest lattice and combined by a weighted sum,

$$\boldsymbol{Z} = \sum_{k=0}^{2} \alpha_k \odot \mathrm{Up}(\boldsymbol{L}_k) \in \mathbb{R}^{256\times H\times W}, \qquad \sum_{k=0}^{2} \alpha_k = 1 \text{ pointwise},$$

where $\odot$ is the Hadamard product. TripletAttention then applies attention across the three rotations of the (channel, height, width) triplet (each branch: a Z-pool that concatenates average- and max-pooled maps, a 7×7 convolution, and a sigmoid gate), and a final 1×1 convolution compresses the fused map to 32 channels,

$$\boldsymbol{Y} = \mathrm{Conv}_{1\times1}(\boldsymbol{Z};\, 32) \in \mathbb{R}^{32\times H\times W}.$$

Coordinate sampling (grid_sample). Each cell's 32-dimensional spatial feature is read from $\boldsymbol{Y}$ at its fractional grid coordinate $g_i \in [-1,1]^2$ by bilinear interpolation:

$$\boldsymbol{r}_i = \mathrm{grid_sample}(\boldsymbol{Y},\, g_i) \in \mathbb{R}^{32}, \qquad g_i = \left(\frac{2\,\tilde{c}_i}{W} - 1,\ \frac{2\,\tilde{r}_i}{H} - 1\right),$$

$$\tilde{c}_i = \frac{x_i - x_{\min}}{b} + 1/2\,, \qquad \tilde{r}_i = \frac{y_i - y_{\min}}{b} + 1/2.$$

This differentiable readout binds each cell's continuous position to the rasterised neighbourhood context and keeps the model permutation-correct with respect to cell order.

External attention and graph encoder. ExternalAttention replaces the $N \times N$ self-attention matrix with a small learned external memory $\boldsymbol{M}_k, \boldsymbol{M}_v \in \mathbb{R}^{S\times d}$:

$$\mathrm{Attn}(\boldsymbol{r}) = \mathrm{softmax}\left(\frac{(\boldsymbol{r}\,\boldsymbol{W}_q)\,\boldsymbol{M}_k^{\top}}{\sqrt{d}}\right)\boldsymbol{M}_v,$$

which is linear in the number of cells and order-invariant. The Fourier graph network (FGN) treats the 32-channel vector as a sequence and returns the 16-dimensional niche embedding, and a single linear decoder reconstructs the input:

$$\boldsymbol{e}_i = \mathrm{FGN}(\boldsymbol{r}_i;\ \mathrm{pre} = 16,\ \mathrm{emb} = 8,\ \mathrm{seq} = 32,\ \mathrm{hid} = 32) \in \mathbb{R}^{16}, \qquad \widehat{\boldsymbol{x}}_i = \boldsymbol{W}_{\mathrm{dec}}\,\boldsymbol{e}_i + \boldsymbol{b}_{\mathrm{dec}} \in \mathbb{R}^{39}.$$

Objective. With $N_{k_0}(i)$ the base-scale ($k_0 = 30$) neighbourhood, the loss summed over the slides is

$$\mathcal{L} = \sum_s \left[\frac{1}{39\,N_s} \sum_{i=1}^{N_s} \| \widehat{\boldsymbol{x}}_i - \boldsymbol{x}_i \|_2^2 + W_{\mathrm{smooth}} \frac{1}{k_0\,N_s} \sum_{i=1}^{N_s} \sum_{j \in N_{k_0}(i)} \| \boldsymbol{e}_i - \boldsymbol{e}_j \|_2^2 \right], \qquad W_{\mathrm{smooth}} = 1.0.$$

The first term forces the 16-dimensional embedding to retain the input through the decoder; the second is an explicit spatial-smoothness prior that pulls the embeddings of neighbouring cells together, so the embedding is a smooth field over the tissue rather than an unconstrained per-cell autoencoder latent.

Clustering and separation. Niches minimise the KMeans inertia over the embeddings, and separation is the mean silhouette:

$$\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{N} \| \boldsymbol{e}_i - \mu_{a(i)} \|_2^2 \quad (K = 9), \qquad \bar{s} = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ the mean distance to the nearest other cluster. Cross-slide label agreement is the adjusted Rand index (ARI) between label sets $\mathcal{A}$ and $\mathcal{B}$,

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{1/2 \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

with n_{ij} the overlap count of class i of $\mathcal{A}$ and class j of $\mathcal{B}$.

Granuloma signature and reversal score. A gene g enters the granuloma signature if it clears an expression floor and a fold-change threshold on its granuloma-versus-alveolar niche contrast δ_g:

$$\delta_g = \frac{1}{|\mathcal{N}_G|}\sum_{n\in\mathcal{N}_G}\bar{E}_{gn} \;-\; \frac{1}{|\mathcal{N}_A|}\sum_{n\in\mathcal{N}_A}\bar{E}_{gn}\,, \qquad g \in \text{sig} \Leftrightarrow \left(\max_n \bar{E}_{gn} \geq 0.5\right) \wedge |\delta_g| \geq 0.5,$$

with $\mathcal{N}_G = \{N4, N6\}$, $\mathcal{N}_A = \{N0, N2\}$, and $\bar{E}_{gn}$ the mean normalised expression of gene g in niche n; the up- and down-subsets are $\{\delta_g \geq 0.5\}$ and $\{\delta_g \leq -0.5\}$, and significance is assessed by a niche-label permutation test. For a perturbation d, the signature-reversal score sums the hypergeometric enrichment of the signature up-set in d's down-set and of the signature down-set in d's up-set:

$$\text{rev_score}(d) = \max_{L\in\mathcal{L}_\downarrow}\left(-\log_{10}P_{\text{hyp}}(S_\uparrow \cap L_\downarrow)\right) \;+\; \max_{L\in\mathcal{L}_\uparrow}\left(-\log_{10}P_{\text{hyp}}(S_\downarrow \cap L_\uparrow)\right),$$

where P_{hyp} is the hypergeometric (Enrichr) enrichment p-value and $\mathcal{L}_\uparrow, \mathcal{L}_\downarrow$ are the up/down gene sets of perturbation d in a given library. A high score means the perturbation drives the granuloma signature in the opposite direction, nominating a candidate therapeutic direction.

Joint training

The encoder is trained once with shared weights across both slides (joint cross-slide training); each training step processes one slide (batch = 1 slide), so the shared encoder sees both slides across steps. Normalisation is z-scoring of the input features, computed either globally (cross-slide pooled statistics; default) or per slide. The loss is MSE(reconstruction, input) + W_smooth · spatial-kNN-smoothing(embedding), where the smoothing term penalises embedding differences between spatial neighbours and W_smooth = 1.0 by default. Training used the Adam optimiser at lr = 1e-3 for 500 steps at the default. Default hyperparameters: seed 0, smooth 1.0, steps 500, lr 1e-3, modality-mask rate 0.3, input Gaussian noise SD 0.1, K = 9, embedding dimension 16, fuse channels 32, 2,000 HVG, 30 PCs, kNN scales {10, 20, 30}. Input augmentation consists of adding Gaussian noise and masking a fraction of the composition/expression blocks; both were active at the default and are shown by the augmentation ablation to be functional.

Niche clustering and markers

Niches were obtained by KMeans on the 16-dimensional embedding with K = 9, applied jointly so that niche identity is shared across slides. The silhouette coefficient of the KMeans partition in the embedding is reported as the separation metric. Marker genes per niche were selected as genes with expression above the 75th percentile of positive values and log2FC > 0 within the niche, ranked by log2 fold change to the top six, after

removing unsymbolised ENSG* identifiers and non-coding/pseudogene names (LINC / -AS / -DT / pseudogene suffixes); candidates were required to be detected in ≥10% of cells. The slide-specific marker lists are given in Supplementary Table S1.

Ligand–receptor analysis

Niche-level cell–cell communication was summarised as a niche-to-niche ligand–receptor network (Fig. 4e), computed from the mean expression of ligand–receptor gene pairs aggregated per niche, in the spirit of established cell–cell-communication inference. This analysis is used as a qualitative consistency check on the niche marker identities.

SP110 hub analysis

SP110 was selected as a candidate hub because it is a known host TB-susceptibility/nuclear-body protein and one of the measured spatial markers used to cross-check the deconvolution. For each cell we computed an **SP110–IFN hub degree** — the number of SP110–IFN^{+} cells among its six spatial nearest neighbours — and an **SP110–IFN module score**, and mapped both across B1 and P1 (Fig. 4f, g). We additionally related the SP110–IFN module score to an unfolded-protein-response (UPR) module score on each slide (Fig. 4g insets).

Granuloma signature derivation

The granuloma signature was derived on slide B1 by contrasting the granuloma-associated niches (N4 and N6) against the alveolar/airway niches (N0 and N2). A gene entered the signature if it satisfied an expression floor of 0.5 and $|\log2FC| \geq 0.5$. This yielded 2,432 detected genes and, after filtering, a final signature of **324 genes (233 up-regulated, 91 down-regulated)** (Fig. 6d). Gene-ontology biological-process (GO-BP) enrichment of the signature is summarised in Fig. 6e. Enrichment was computed with clusterProfiler against org.Hs.eg.db with Benjamini–Hochberg adjustment; the leading terms and their statistics are reported in Results.

Drug repurposing

The 324-gene signature was queried against a **merged perturbation library** — the LINCS L1000 chemical-perturbation set, GEO perturbation experiments, NIBR DRUGseq 2025, and the original CMAP, combined (36,713 perturbation records; 36,001 unique after removing sample/batch suffixes and de-duplicating by primary drug name) — with a signature-reversal score. For each perturbation, the score aggregates the reversal of the up- and down-regulated signature sides (rev_up, rev_dn); a high score indicates that the perturbation

shifts the granuloma signature in the opposite direction, nominating a potential therapeutic direction. Hits were ranked by reversal score (Supplementary Table S2; Fig. 6f, g).

Molecular docking

A focused six-ligand panel (withaferin A, nintedanib, dasatinib, niclosamide, JQ1, vorinostat) was docked against the host target SP110 using AutoDock Vina 1.2.5 (standalone) with exhaustiveness 32 and 20 output modes per ligand. Best-mode affinities are reported in kcal/mol (Fig. 6h); docking reproducibility was assessed by the number of converged modes and the mode-1 RMSD between repeat sub-samplings. Docking is reported as in-silico binding feasibility only (Supplementary Table S3).

Robustness, ablation, and sensitivity protocols

Robustness. Input noise tolerance was measured by adding Gaussian noise of scale x per feature SD to the normalised input, re-running the full pipeline, and computing the adjusted Rand index (ARI) between the perturbed niche labels and the clean reference labels (ARI ≡ 1 at x = 0 by construction); noise scales were 0.25, 0.50, 0.75, and 1.00 (Fig. 5f). Missing-modality tolerance was measured by zeroing the 7-dimensional composition block or the 30-dimensional expression-PC block and computing ARI against the clean labels. Cross-slide transfer was measured by leave-one-dataset-out (LODO): an encoder trained on the single other slide embeds the held-out slide, and its niche labels are compared (ARI) against the jointly trained reference labels.

Ablation. Each module (or the multi-scale construction) was removed in turn, and each ablation was run with seeds 0/1/2. Δ was computed as (ablation − full model) paired by seed, and reported as mean ± sd. A Δ was treated as stable only if all three seeds shared a sign and |mean| > sd (marked † in Table 5; Fig. 5c).

Sensitivity. A single-factor sweep varied one hyperparameter at a time while holding the rest at the default: smooth ∈ {0.5, 1.0, 2.0, 4.0, 8.0}, lr ∈ {1e-4, 2e-4, 5e-4, 1e-3, 2e-3}, mask ∈ {0, 0.1, 0.3, 0.5, 0.7}, noise ∈ {0, 0.1, 0.2, 0.5}, steps ∈ {250, 500, 1000}, plus combination and 2-D knee configurations (34 configurations total; Fig. 5d, e). The default configuration was included as a reference point. Lead candidates were re-run with three seeds; a candidate was only considered a valid replacement for the default if it retained robustness within the seed noise band (Table 6).

Normalisation comparison. Global (cross-slide pooled) versus per-slide normalisation was compared in a 3 × 2 paired design across seeds 0/1/2 for eight metrics; Δ(global − per-slide) was reported as mean ± sd, and a difference was considered seed-stable only if |mean| > sd. Seven of eight metrics failed this criterion (Table 6c; Fig. 6a).

Computational cost

Parameter counts were obtained by instantiating MuSpec(c_in = 39) and summing numel() over all parameters. Throughput was timed with synthetic inputs matching the real shapes (three (1, 39, 224, 224) rasters plus a per-cell grid at N = 2,000), after 5 warm-up iterations, over 100 timed iterations with torch.cuda.synchronize(); the reported value is the mean ms per iteration. Peak memory was read from torch.cuda.max_memory_allocated() for the isolated step and from nvidia-smi during the full B1 + P1 pipeline. All experiments ran on a single RTX 4080 SUPER (32 GB).

Results

The Results are organised around six figures: the model and its cost (Fig. 1), the joint niche map (Fig. 2), biological characterisation of the niches (Fig. 3), cross-slide conservation and the SP110–IFN hub (Fig. 4), benchmarking and robustness (Fig. 5), and translation (Fig. 6).

MuSpec architecture and computational cost

MuSpec is a multi-scale encoder whose design follows a single principle: represent each cell not only by its own state but by its state at several spatial neighbourhood scales, and fuse those scales before embedding (Fig. 1). The forward pass (Table 2a) begins with the 39-dimensional per-cell input, applies kNN smoothing at three scales (k = 10, 20, 30), rasterises each scale onto a $\leq 224^2$ grid, and projects each raster with a 1×1 convolution (to 512, 256, and 256 channels). Average-pooling aligns the three scales (×4 and ×2), after which **ASFF** performs adaptive spatial feature fusion of the three levels into a single 256-channel map. **TripletAttention(256)** then applies spatial attention across the (H, W, C) dimensions. A 1×1 convolution compresses the map to 32 channels, and each cell's embedding is sampled from this map via grid_sample at its fractional coordinate. **ExternalAttention(32, 8 heads)** applies channel-wise self-attention with a learned external memory, and **FGN**, a Fourier graph network instantiated as FGN(pre_length=16, embed_size=8, feature_size=1, seq_length=32, hidden_size=32), produces the final 16-dimensional niche embedding. A linear

decoder reconstructs the 39-dimensional input. The loss is MSE(reconstruction, input) + W_smooth · kNN-smoothing(embedding), and the encoder is trained jointly across both slides with shared weights. The architecture schematic (Fig. 1) is drawn from these components, with boxes distinguishing our pipeline code from the four attention-and-fusion modules; the schematic also depicts a linear read-out head that is not used in the unsupervised losses of this study (see Methods, MuSpec architecture).

The four functional modules in this pipeline implement published attention-and-fusion architectures (Table 2b): ASFF (adaptive spatial feature fusion), TripletAttention (cross-dimension attention), ExternalAttention (external-memory attention), and FGN (Fourier graph network). The whole pipeline — data loading, the composition block, the three-scale kNN smoothing, rasterisation, the 1×1 projections, the pooling pyramid, the fuse head, the decoder, the joint-training loop, the normaliser, the augmentation (Gaussian noise plus modality masking), and the robustness harness — was written for this study. We state this provenance plainly because it bounds the novelty claim: the methodological contributions of this paper are the data representation, the multi-scale wiring, the joint cross-slide training, and the benchmark protocol, not the individual attention or fusion modules.

Table 2. MuSpec architecture and module provenance.

2a. Forward pass (batch = 1 slide; *N* = cells; c_in = 39; grid ≤ 224²).

#	Stage	Operation	Module / source	Shape
0	Input	7 composition ‖ 30 expr-PC ‖ 2 (x,y)	assembly	(N,39)
1	Multi-scale graph	kNN smoothing k = 10/20/30	assembly (kNN)	(N,39)×3
2	Rasterise	bin-average → ≤224²	assembly	(1,39,H,W)×3
3	Scale projection	1×1 conv → 512 / → 256 / → 256	nn.Conv2d	(1,·,H,W)×3
4	Pyramid	avg-pool ×4 / ×2	F.avg_pool2d	l0 / l1 / l2
5	Cross-scale fusion	ASFF(level=2)	this study	(1,256,·,·)
6	Spatial attention	TripletAttention(256)	this study	(1,256,·,·)
7	Fuse head	1×1 conv 256 → 32	nn.Conv2d	(1,32,H,W)
8	Sampling	grid_sample at cell coords	F.grid_sample	(N,32)
9	Channel attention	ExternalAttention(32, 8 heads)	this study	(N,32)
10	Graph encoder	FGN(pre16, emb8, feat1, seq32, hid32)	this study	(N,16) ← niche embedding
11	Decoder	Linear 16 → 39	nn.Linear	(N,39)

Loss = MSE(rec, X) + W_smooth · kNN-smooth(emb); the encoder is trained jointly across slides (shared weights); clustering is KMeans (K = 9) on the 16-dimensional embedding.

2b. Module provenance — the four functional modules implement published attention-and-fusion architectures; both their implementations and the surrounding multi-scale pipeline were written for this study.

MuSpec is small and cheap (Table 7). The model has **782,251 trainable parameters**, of which ASFF accounts for 734,451 (93.9%) — the fusion, not the encoder, dominates the parameter budget. Training throughput was 36.9 ms per iteration (27.1 iterations/s) at N = 2,000 cells per step (forward + backward + Adam), peak GPU memory was 1.0 GB for an isolated step and ≈7.0 GB end-to-end on a 32-GB card, and convergence was reached by **500 steps**. A single configuration trains in ≈1–2 minutes, and the full 34-configuration sweep took ≈1–2 hours on one RTX 4080 SUPER. Every robustness, ablation, and sensitivity experiment in this paper therefore ran on a single consumer workstation, which is what makes the three-seed and 34-configuration studies below practically possible.

Table 7. Model size and computational cost.

Quantity	Value
Total trainable parameters	**782,251** (ASFF 734,451 = 93.9%)
Component breakdown	p512 20,480 · p256 10,240 · fuse 8,224 · FGN 7,464 · decoder 663 · ExtAttn 512 · TripAttn 200 · cls 17
Train throughput	**36.9 ms/iter · 27.1 iter/s** (N = 2,000, fwd + bwd + Adam)
Peak GPU memory (isolated / end-to-end)	**1.0 GB / ≈7.0 GB** (RTX 4080 SUPER 32 GB)
Steps to convergence	**500**
Per-config / full 34-config sweep	**≈1–2 min / ≈1–2 h** (single card)

MuSpec recovers conserved spatial niches on a granuloma and its peri-lesional margin

We analysed two Stereo-seq cellbin slides from human TB lung tissue (Table 1). Slide **B1** (Y40012NC13) captures a pulmonary **granuloma** (36,382 cells); slide **P1** (Y40012LA11) captures **peri-lesional** lung tissue from the same lung (51,834 cells). For each cell we constructed a 39-dimensional feature vector: a 7-dimensional **composition block** (the estimated fractions of T cells, dendritic cells, NK cells, $CD8^+$ T cells, B cells, $CD4^+$ T cells, and monocytes, obtained by cell2location deconvolution against a 146,345-cell peripheral-blood scRNA-seq reference), a 30-dimensional **expression-PC block** (raw counts → CP10K → log1p → 2,000 highly variable genes → PCA with 30 components), and 2 spatial coordinates. The composition block and the expression-PC block are the two modalities probed in the missing-modality experiments below; the spatial coordinates anchor each cell to its grid position.

MuSpec produced a 16-dimensional niche embedding per cell, and KMeans with K = 9 partitioned each slide into nine **joint niches** (shared across slides by construction of the jointly trained encoder; Fig. 2a, b).

Qualitatively, the two slides differ in the way the biology predicts: B1 domains are smaller and more fragmented, consistent with the high transcriptional heterogeneity of granuloma tissue, whereas the P1 domains occupy larger, more contiguous territories consistent with peri-lesional lung. The per-niche cell-type composition (Fig. 2c, d) and its compact sunburst summary against the blood-reference classes (Fig. 2e) show that the nine niches are non-redundant mixtures rather than single-cell-type clusters, which is the intended behaviour of a niche model: a niche is a local microenvironment, not a cell type.

Table 1. Datasets and reference cohort.

#	Role	Platform	Sample ID	Region / tissue	Cells (n)	Feature space	Role in this study
B1	Discovery	Stereo-seq (cellbin)	Y40012NC13	pulmonary **granuloma**	**36,382**	39-dim = 7 composition ‖ 30 expr-PC ‖ 2 (x,y)	niche discovery; robustness (noise / missing-modality)
P1	Discovery	Stereo-seq (cellbin)	Y40012LA11	**peri-lesional** lung	**51,834**	39-dim = 7 composition ‖ 30 expr-PC ‖ 2 (x,y)	cross-slide validation; LODO target
REF	Deconvolution ref	10x scRNA-seq	5 HC / 5 DS / 5 DR	peripheral blood	**146,345**	7 cell-type classes (composition prior)	cell2location → per-spot 7-dim composition

Expression pipeline: raw counts → CP10K → log1p → 2,000 HVG → PCA(30). The 7 reference classes are T, dendritic, NK, $CD8^+$ T, B, $CD4^+$ T, and monocyte. The two slides are identified by their BGI sample IDs (B1 = Y40012NC13; P1 = Y40012LA11; see Data availability). Cell counts are the intersection of deconvolution-scored barcodes, coordinates, and expression, and differ by ±2 from the raw cellbin totals (36,384 / 51,835); the spatial panels in Fig. 2 show the N = 33,769 (B1) and N = 48,141 (P1) cells at which niche labels and effective coordinates coincide.

Biological characterisation of the joint niches

To test whether the learned niches are biologically meaningful rather than geometrically convenient, we computed per-niche marker genes (genes expressed above the 75th percentile of positive values and enriched in the niche at log2FC > 0; top six per niche; Supplementary Table S1; Fig. 3a, d). The joint niches carry coherent, non-overlapping biological identities (Fig. 3). The richest and most translationally relevant signals were on the peri-lesional slide P1: niche **N1** was marked by **SPP1, MMP9, PLA2G2D, ADAMDEC1, CCL19, CTSK** — an $SPP1^+$ macrophage, osteoclast-like (CTSK) and chemokine-rich myeloid signature characteristic of the

granuloma core; niche **N6** by **SFRP2, GREM1, INHBA, THBS2, COL1A1, IGLC3** — a fibrotic, myofibroblast/ECM-deposition signature consistent with the fibrotic ring that surrounds granulomas; niche **N5** by **MUC5B, GRP, C20orf85, BPIFB1, SCGB1A1, SAA2** — an airway/club-cell signature, indicating that peri-lesional tissue retains patent airway structure; and niche **N7** by **IL11, ITLN1, PLA2G2A, PI16, CLDN1, TFPI2** — a matrix/fibroblast signature centred on the IL-11 fibrotic axis. On the granuloma slide B1 the same joint niches showed partly different top markers — for example B1 N7 carried **MS4A1** (B cells) and B1 N4 carried **FOLR3/TIMD4** (tissue macrophages) — which we read as B1 being transcriptionally more dispersed, so that individual markers are less specific, while the shared axes (myeloid–fibroblast, airway) persist. The clearest cross-slide anchor is niche **N5**, which retains **SCGB1A1** on both slides.

The niche identities are reinforced by two orthogonal composition views. A niche × cell-type composition heatmap (Fig. 3b, e) shows that each niche is dominated by a characteristic cell-type mixture, and a niche × blood-reference-class score heatmap (Fig. 3c, f) anchors those mixtures to the deconvolution reference classes, so that myeloid niches score high on the monocyte/dendritic axis and lymphoid niches on the B/NK axis. A niche–cell-type association network (Fig. 3g) places the granuloma and peri-lesional niches in a single graph and shows that individual cell types connect preferentially to particular niches rather than to all niches equally.

Two features of this result matter for the paper. First, the niches recovered are microenvironments with mixed composition, exactly as a niche model should produce, and their markers are interpretable without any supervision or manual annotation. Second, the cross-slide overlap (shared N5 airway identity; a shared myeloid–fibroblast axis) is direct evidence that joint niche discovery produces a structure that transfers across the two slides instead of overfitting one.

Cross-slide conservation, niche communication, and the SP110–IFN hub

Having established interpretable niches, we asked how the niche composition is conserved or reallocated between the granuloma and the peri-lesional margin (Fig. 4). The per-niche shares of the two slides (Fig. 4a) show that some niches occupy similar proportions on both slides (the airway N5 is the clearest example) while others are strongly slide-specific, consistent with the granuloma-versus-margin contrast that motivated the two-slide design. A niche-signature correlation between B1 and P1 (Fig. 4b) separates a shared set of niche signatures (labelled B1+P1_both) from slide-specific ones (B1_only, P1_only), and an alluvial (ribbon) diagram (Fig. 4c) together with a faceted stacked-area view of composition proportions (Fig. 4d) summarises

the same reallocation. We then built a niche-level ligand–receptor map (Fig. 4e), which produced a network of niche-to-niche signalling consistent with the marker identities — for example, signalling converging on the myeloid/granuloma niche and on the fibroblast niche. We report this map as a qualitative consistency check on the niche identities rather than as a quantitatively validated signalling hierarchy.

Among the genes that recurred in the granuloma-side signal, **SP110** stood out as a candidate hub: SP110 is a known host TB-susceptibility/nuclear-body protein and was also one of the measured spatial markers used to cross-check the deconvolution. We therefore asked whether SP110 occupies a central position in the niche network (Fig. 4f, g). We computed a per-cell **SP110–IFN hub degree** — the number of SP110–IFN^{+} cells among a cell's six spatial nearest neighbours — and mapped it across B1 and P1 (Fig. 4f), and we mapped an SP110–IFN module score on both slides (Fig. 4g). The consistent picture is that SP110 associates with the myeloid/granuloma end of the tissue, in line with its role in macrophage-intrinsic host defence. Across slides, the SP110–IFN module score showed no consistent relationship with an unfolded-protein-response (UPR) module score (B1: $R = 0.01$, $P = 9.73 \times 10^{-1}$; P1: $R = -0.26$, $P = 5.07 \times 10^{-1}$; Fig. 4g insets). Because this analysis is descriptive and rests on two slides and a deconvolution-derived composition block, we treat the "SP110 hub" as a hypothesis to be tested orthogonally, not as an established hub.

Matched-K benchmarking, module ablation, and robustness

Matched-K separation versus BANKSY. To place MuSpec's niche separation on an external reference scale, we compared it with **BANKSY** (10), a strong geometry-based baseline, using a deliberate **matched-K** protocol: for each setting of BANKSY's resolution parameter we obtained a specific number of spatial domains *K*, then re-ran KMeans on the MuSpec and single-scale-base embeddings at the *same K* and computed silhouette in the 16-dimensional embedding. This design follows the matched-comparison discipline exemplified by scPerturBench (15) and avoids the unfair framing in which each method is evaluated at its own preferred number of clusters.

Across all eight matched-K conditions, MuSpec separated niches far more sharply than BANKSY (Table 4, Fig. 4h–j). On B1 the silhouette difference Δ(MuSpec − BANKSY) ranged from +0.315 (K = 14) to +0.374 (K = 30); on P1 it ranged from +0.314 (K = 21) to +0.398 (K = 46). BANKSY's own silhouette was near zero or negative at higher resolution (e.g. −0.052 on B1 at K = 30 and −0.126 on P1 at K = 46). We are careful about the interpretation of this gap: BANKSY is a geometric segmentation method and does not optimise an

embedding silhouette, so the correct statement is not that BANKSY lacks discriminative power but that **at a given number of domains, the MuSpec embedding yields far more compact niche separation**. MuSpec also slightly outperformed its own single-scale base at every matched K (Δ +0.017 to +0.043), with a consistent sign; because the magnitude lies within the seed noise band, we treat this as suggestive and rely on the three-seed single-scale ablation (below) for the multi-scale claim. MuSpec's own K = 9 silhouette was 0.322 on B1 and 0.285 on P1.

The agreement between MuSpec (K = 9) labels and BANKSY domains across the matched-K conditions was moderate: ARI 0.196–0.228 and NMI 0.341–0.369 on B1, and ARI 0.062–0.090 and NMI 0.145–0.150 on P1. Moderate rather than high agreement is the informative outcome: it shows that MuSpec and BANKSY capture partially overlapping but non-equivalent structure, which supports MuSpec's positioning as a source of incremental niche information rather than a re-derivation of geometric domains. The qualitative counterpart of these numbers is the BANKSY domain map itself (Fig. 5a, b): at λ = 0.8, k_geom = 30 and resolution = 0.8, BANKSY fragments B1 into 25 groups and P1 into 37 groups, i.e. many small geometric domains rather than the nine joint niches of Fig. 2.

Table 4. Matched-K benchmark (MuSpec vs BANKSY vs single-scale base).

Slide	K	sil BANKSY	sil MuSpec@K	sil single-scale base@K	Δ(MuSpec−BANKSY)	ARI (K=9 vs BANKSY)	NMI
B1	14	0.021	0.336	0.309	**+0.315**	0.204	0.341
B1	17	0.011	0.329	0.291	**+0.318**	0.228	0.369
B1	25	−0.016	0.316	0.299	**+0.332**	0.225	0.360
B1	30	−0.052	0.322	0.291	**+0.374**	0.196	0.350
P1	21	−0.031	0.283	0.247	**+0.314**	0.090	0.150
P1	26	−0.079	0.273	0.238	**+0.352**	0.077	0.145
P1	37	−0.124	0.271	0.241	**+0.395**	0.064	0.145
P1	46	−0.126	0.272	0.229	**+0.398**	0.062	0.150

MuSpec K = 9 silhouette: **B1 0.322 / P1 0.285**. Δ(MuSpec − single-scale base) ranged +0.017 to +0.043 across matched K. N_common: B1 33,769; P1 48,141.

Module ablation (three-seed). A single-seed ablation is not trustworthy on this data, so every ablation was run with **three seeds (0/1/2)**, with Δ computed as (module removed) − (full model) and paired by seed, reported as mean ± sd (Table 5, Fig. 5c). We define a result as stable only when all three seeds share a sign and |mean| > sd. This discipline changed a conclusion: a single-seed run had suggested that dropping the Fourier graph network (FGN) sharply increased separation (+0.097), but across three seeds the effect collapsed to +0.042 ± 0.077

(seed-wise +0.097 / −0.047 / +0.075). That finding is therefore **not** reported as a result; it is recorded here only as an example of why seed control is essential.

The three-seed ablation supports one strong, stable conclusion and several weak-to-mid ones. The strong result is that **ExternalAttention implements a separation↔robustness trade-off**: removing it reduced niche separation (Δsilhouette −0.093 ± 0.030, stable) while raising robustness and transfer (Δnoise-robustness +0.137 ± 0.104, Δdrop-composition P1 +0.272 ± 0.198, Δ(B1→P1) +0.218 ± 0.057, Δ(P1→B1) +0.362 ± 0.096 — 5 of 6 metrics stable). This is the cleanest module-level evidence for the frontier that the hyperparameter sweep independently revealed. The remaining modules are weaker contributors in the robustness direction: dropping FGN reduced noise-robustness (Δ −0.101 ± 0.090, stable), and dropping ASFF or TripletAttention reduced drop-composition robustness on B1 (Δ −0.097 ± 0.022 and −0.125 ± 0.109, respectively). Replacing the multi-scale construction with a single scale reduced separation (Δsilhouette −0.040 ± 0.031, stable), which is the same direction as the matched-K "MuSpec beats single-scale base" result and is its seed-controlled confirmation. Overall the ablation says that the architecture is an integrated whole — **no module is removable losslessly** — while the only individually strong and seed-stable statement is the ExternalAttention trade-off.

Table 5. Module ablation, three seeds (Δ = drop − full, paired by seed). Default absolute values: si$\bar{l}$ 0.332 ± 0.030; n̄1.0 0.654 ± 0.095; dcB1 0.658 ± 0.073; dcP1 0.490 ± 0.108; B1→P1 0.162 ± 0.058; P1→B1 0.226 ± 0.118. † = same sign across all three seeds and |mean| > sd.

Module removed	Δsi$\bar{l}$	Δn̄1.0	Δdc(B1)	Δdc(P1)	Δ(B1→P1)	Δ(P1→B1)
ASFF	**−0.014** †	−0.009	**−0.097** †	+0.015	+0.010	**+0.126** †
FGN	+0.042	**−0.101** †	−0.076	−0.039	−0.032	−0.023
ExternalAttention	**−0.093** †	**+0.137** †	+0.049	**+0.272** †	**+0.218** †	**+0.362** †
TripletAttention	+0.002	−0.029	**−0.125** †	+0.004	+0.000	+0.057
single-scale (multi → 1)	**−0.040** †	+0.120	−0.067	+0.021	+0.050	**+0.238** †

Robustness and hyperparameter sensitivity. Two robustness questions are addressed here, both with the seed-controlled discipline used for the ablation: how the model behaves under input perturbations, and how sensitive it is to hyperparameters.

Normalisation. We compared **global** (cross-slide pooled) normalisation with **per-slide** normalisation. In a single-seed comparison (Table 3), the global setting looked consistently better: it gave higher noise-robustness

(ARI@1.0 = 0.806 vs 0.793 on B1 and 0.711 vs 0.648 on P1), better missing-modality tolerance (drop-composition 0.683 vs 0.609 on B1 and 0.518 vs 0.421 on P1), and better cross-slide transfer (LODO 0.187 vs 0.153 on B1 and 0.351 vs 0.298 on P1), at essentially equal silhouette. It would be tempting to write "global normalisation improves robustness". The three-seed paired design does not support that claim. Of eight metrics, **seven changed sign across seeds** (Table 6c, Fig. 6a), and the mean Δ was far smaller than its seed standard deviation; the only stable difference was B1 silhouette, which favoured per-slide normalisation by a very small 0.010 ± 0.003. We therefore report the neutral and honest result: **shared versus per-slide normalisation produces no seed-stable difference** in cross-slide generalisation or damage tolerance, with only a marginal separation advantage to per-slide normalisation on B1. The robustness claims in this paper rest on the multi-scale and permutation statistics described below, not on this normalisation comparison.

Hyperparameter sensitivity and the separation↔robustness frontier. A single-factor sweep of 34 configurations (Table 6, Fig. 5d–f) produced a clear and monotone **separation–robustness frontier**: configurations that raised niche separation paid for it in robustness. Lowering the learning rate (1e-4 or 2e-4) raised mean silhouette to 0.539 / 0.531 but dropped P1 noise-robustness to 0.445 / 0.448 and drop-composition accuracy to 0.327–0.461; higher smoothing behaved similarly. The default configuration (smooth 1.0, lr 1e-3, steps 500, mask 0.3, noise 0.1) sat at the **robustness end** of the frontier and was the robustness optimum in the entire sweep (mean noise-robustness 0.758; drop-composition 0.683 on B1 and 0.518 on P1, both the highest observed). A leading high-separation candidate (smooth 4.0 + lr 5e-4, three seeds) reached a higher mean silhouette (0.458 ± 0.054) but with badly unstable P1 robustness (P1 noise-robustness 0.466 ± 0.226 and P1 drop-composition 0.282 ± 0.190), so we do not recommend it as a replacement for the default. The sweep also showed flat sensitivity with **no cliffs** — neighbouring settings varied by about ±0.06–0.14, the same order as the seed standard deviation — which is a positive reproducibility result: the model does not sit on a knife edge. Finally, 500 steps is a **selected training budget**, not a convergence point: steps 250/500/1000 gave mean silhouettes 0.358/0.314/0.290 and noise-robustness peaked at 500 steps (0.758), so beyond 500 steps silhouette does not increase and robustness declines. We adopt 500 steps as the default because it is where robustness is maximal and further training adds cost without benefit (Table 7); 500 is therefore a robustness-optimal budget, not a separation-optimal or "converged" one. An augmentation ablation provided direct positive evidence that the augmentation is functional rather than cosmetic: the default augmentation gave the highest robustness in the sweep, and removing the modality mask was the most damaging single change (mean noise-robustness 0.758

→ 0.556), followed by removing noise (→0.686) and removing both (→0.709). The noise-robustness curves themselves (Fig. 5f) make the same point directly: ARI against the clean labels decays smoothly with input-noise SD, and the global and per-slide normalisation variants are almost indistinguishable at every noise level on both slides.

Robustness–covariate summary. A summary view of the robustness relationships (Fig. 6b, c) relates a per-configuration covariate to the six robustness metrics; the relationship is negative for all six metrics (Spearman r from −0.22 to −0.65; q from 4.4×10^{-5} to 0.21), i.e. higher values of the covariate accompany lower robustness. The covariate is the **mean silhouette separation** of the configuration, $\bar{s} = 1/2\,(s_{\mathrm{B1}} + s_{\mathrm{P1}})$, the average of the B1 and P1 KMeans silhouette at K = 9 (abbreviated *MTAP* in the figure labels, a leftover variable name from the plotting template); it is a per-configuration scalar and so takes one value per configuration of the sweep. The correlation bubble heatmap (Fig. 6c) shows the same six relationships with their −log10 P values; the single non-significant relationship (q = 0.21) is reported as such rather than interpreted.

Table 3. Cross-slide robustness under global vs per-slide normalisation (single-seed reference values; see text and Fig. 6a for the three-seed verdict).

Slide	Norm	n	Silhouette ↑	ARI@0.25	ARI@0.50	ARI@0.75	ARI@1.00	drop comp	drop PC	LODO ARI
B1 granuloma	global	36,382	0.349	0.951	0.896	0.850	**0.806**	**0.683**	0.400	**0.187**
B1 granuloma	per-slide	36,382	**0.355**	0.941	0.879	0.840	0.793	0.609	**0.401**	0.153
P1 peri-lesion	global	51,834	**0.280**	0.916	0.839	0.769	**0.711**	**0.518**	**0.386**	**0.351**
P1 peri-lesion	per-slide	51,834	0.262	0.910	**0.859**	**0.780**	0.648	0.421	0.349	0.298

Definitions: silhouette = mean silhouette of KMeans(K = 9) in the 16-dimensional embedding; ARI@noise *x* = adjusted Rand index versus clean labels after adding Gaussian noise of scale *x* per-feature SD; drop comp / drop PC = ARI after zeroing the 7-dimensional composition block / 30-dimensional expression-PC block; LODO transfer ARI = leave-one-dataset-out transfer against joint-training labels. global = cross-slide pooled z-scoring; per-slide = within-slide z-scoring.

Table 6. Hyperparameter sensitivity (34 single-factor configurations; condensed), and the three-seed verdict on normalisation.

Group	Representative config	siĪ	n̄1.0	dc (B1 / P1)
—	**default**	0.314	**0.758**	**0.683 / 0.518**
lr	lr 1e-4 / 2e-4	**0.539 / 0.531**	0.551 / 0.574	0.417/0.327 · 0.461/0.395
smooth	2.0 / 4.0 / 8.0	0.320 / 0.359 / 0.335	0.566 / 0.602 / 0.614	↓
mask	mask 0 (off)	0.323	0.556	0.587/0.479
combo	smooth 4.0 + lr 5e-4 (3 seeds)	**0.458 ± 0.054**	0.756 ± 0.089	**0.535 ± 0.081 / 0.282 ± 0.190**

6c. Three-seed paired Δ(global − per-slide). Seven of eight metrics flipped sign across seeds (e.g. P1 noise-robustness Δ = +0.063 / +0.060 / −0.193; B1 drop-composition Δ = +0.074 / −0.040 / +0.183). The only stable difference was B1 silhouette: **−0.010 ± 0.003** (per-slide marginally higher). Verdict: no seed-stable difference between global and per-slide normalisation.

Translation: granuloma signature, host-directed repurposing, and docking

We next asked whether the joint niches support a translational hypothesis. Contrasting the granuloma-associated niches (N4/N6) against the alveolar/airway niches (N0/N2) on slide B1 yielded a **324-gene granuloma signature (233 up, 91 down)** after requiring an expression floor of 0.5 and $|\log2FC| \geq 0.5$ (2,432 genes detected; Fig. 6d). The signature was enriched for fibrosis/ECM, myeloid/antigen-presentation and hypoxia-response biology (Fig. 6e). The leading GO-BP terms were cytoplasmic translation (56 genes; 16.1-fold enrichment; adjusted $P = 2.5 \times 10^{-48}$) and ribosome biogenesis (26 genes; 4.9-fold; adjusted $P = 6.2 \times 10^{-9}$), alongside fibrosis/ECM terms (extracellular matrix organisation; 20 genes; 4.4-fold; adjusted $P = 3.9 \times 10^{-6}$), myeloid/antigen-presentation biology (MHC class II antigen processing and presentation; 8 genes; 14.5-fold; adjusted $P = 6.0 \times 10^{-6}$), and hypoxia/immune-response terms (response to hypoxia; 22 genes; 4.9-fold; adjusted $P = 1.8 \times 10^{-7}$; humoral immune response; 22 genes; 4.6-fold; adjusted $P = 5.2 \times 10^{-7}$). We then queried a merged perturbation library (LINCS L1000 chemical perturbations, GEO perturbations, NIBR DRUGseq 2025, and the original CMAP; 36,713 perturbation records; 36,001 unique after de-duplication) with a signature-reversal score, in which a high score means the perturbation pushes the granuloma signature in the opposite direction (a potential therapeutic direction; Fig. 6f, g).

The top hits were dominated by a mechanistically coherent class: **PPARγ agonists occupied three of the top eight ranks** — pioglitazone (90.08), troglitazone (79.72), and rosiglitazone (63.88) — alongside rituximab (85.36), testosterone (82.74), metoprolol (65.38), motexafin gadolinium (65.10), adenosine triphosphate (64.08), decitabine (61.53), and atorvastatin (58.77) (Supplementary Table S2). Two cross-checks increase our confidence that the signature captures real biology rather than artefact: the anti-CD20 antibody rituximab ranks

second, consistent with the B-cell marker MS4A1 in B1 niche N7; and the PPARγ triad is exactly the class that host-directed-therapy literature associates with dampening excessive inflammation and promoting antimicrobial autophagy in TB. The remaining hits (metoprolol, atorvastatin, estradiol, dexamethasone, vitamin A) are marketed drugs that could enter a repurposing candidate pool directly. Importantly, the reversal score was driven almost entirely by the up-regulated side of the signature (rev_up ≫ rev_dn), so the hits should be read as "reverses the granuloma up-signature", nothing stronger.

To add a binding-level (not efficacy-level) line of evidence, we docked a focused six-ligand panel against **SP110**, the host TB-susceptibility protein that also emerged as the hub candidate above, using AutoDock Vina 1.2.5 (exhaustiveness 32; Supplementary Table S3, Fig. 6h). The strongest affinities were withaferin A (−7.18 kcal/mol) and nintedanib (−6.92 kcal/mol), followed by dasatinib (−6.41), niclosamide (−5.62), JQ1 (−5.55), and vorinostat (−4.54). All six ligands returned 20 converged modes with mode-1 RMSD 0 (repeat sub-sampling consistent), indicating a reproducible docking run. Two of the strongest ligands are mechanistically apt: withaferin A has reported anti-TB/anti-inflammatory activity, and nintedanib is an approved anti-fibrotic drug that matches the fibrotic N6 niche (SFRP2/COL1A1). We stress that docking against SP110 is **in-silico binding-feasibility evidence only**; it is not evidence of activity or efficacy, and it does not establish SP110 as the mechanism of the repurposing hits. The PPARγ triad was not included in this docking panel, so the repurposing and docking results are complementary, not mechanistically linked, at this stage.

Supplementary Table S2. Top-20 signature-reversal hits (merged multi-library screen).

Rank	Drug	rev_score	Rank	Drug	rev_score
1	pioglitazone	90.08	11	estradiol	55.87
2	rituximab	85.36	12	TCDD	54.49
3	testosterone	82.74	13	apratoxin A	54.43
4	troglitazone	79.72	14	hypochlorous acid	54.31
5	metoprolol	65.38	15	Mehp	52.75
6	motexafin gadolinium	65.10	16	tibolone	51.98
7	adenosine triphosphate	64.08	17	vitamin A	51.92
8	rosiglitazone	63.88	18	paclitaxel	51.81
9	decitabine	61.53	19	bisphenol	51.68
10	atorvastatin	58.77	20	dexamethasone	51.27

Supplementary Table S3. Docking of a focused panel against SP110 (AutoDock Vina 1.2.5).

Rank	Ligand	Best affinity (kcal/mol)	n_modes	Note
1	withaferin A	**−7.18**	20	natural product; reported anti-TB/anti-inflammatory
2	nintedanib	**−6.92**	20	approved anti-

Rank	Ligand	Best affinity (kcal/mol)	n_modes	Note
				fibrotic; matches N6 fibrotic niche
3	dasatinib	−6.41	20	kinase inhibitor
4	niclosamide	−5.62	20	marketed anthelmintic; repurposing candidate
5	JQ1	−5.55	20	BET inhibitor
6	vorinostat	−4.54	20	HDAC inhibitor

Discussion

MuSpec gives an unsupervised, computationally light route to spatially resolved niche atlases, and applying it to two single-cell-resolution Stereo-seq slides of human TB lung recovered a set of nine joint niches that are biologically interpretable and partly conserved across a granuloma and its peri-lesional margin. The central advance of the paper is not a new attention mechanism — the four functional modules implement published attention-and-fusion architectures — but the combination of a multi-scale, two-modality cell representation, joint cross-slide training, and a benchmark and stress-test protocol that reports the model's behaviour honestly, including where the evidence is neutral or negative.

What is new, precisely. Because MuSpec builds on published attention-and-fusion designs rather than introducing a new operator, we state the novelty at the level of the system and the protocol, and we enumerate it so that each item can be judged on its own. **(1) Representation.** The per-cell token is a 39-dimensional, multi-scale, two-modality vector — 7-dimensional composition ‖ 30 expression PCs ‖ 2 coordinates — smoothed at three neighbourhood scales, so a cell is described by its local microenvironment at three radii at once. This is what allows a single embedding to act as a *niche* descriptor rather than a cell-type descriptor: the nine recovered clusters are mixed-composition microenvironments (Figs. 2, 3) precisely because the representation is neighbourhood-averaged and multi-scale, and a single-scale autoencoder on the same input loses separation (Δsilhouette −0.040 ± 0.031, seed-stable; Table 5). **(2) Wiring.** The three scales are rasterised to a shared lattice and adaptively fused by ASFF — with learned, location-specific weights — before attention, so scale interaction happens in image space rather than by concatenation; the differentiable grid_sample readout then returns each cell to its fractional coordinate, keeping the pipeline end-to-end differentiable while remaining invariant to cell ordering. **(3) Training.** A single encoder is trained jointly across the granuloma and its peri-lesional margin with a loss that adds an explicit kNN-smoothness prior on the embedding; niche identity is

therefore shared across slides by construction, and the leave-one-dataset-out protocol tests that sharing directly rather than assuming it. **(4) Protocol.** The contribution we consider most useful is negative-but-decisive: matched-K benchmarking, three-seed ablation, and a 34-configuration sweep together reveal a separation↔robustness frontier (Figs. 5, 6) that a single-seed, single-configuration study would have hidden, and they recover a seed-stable ExternalAttention trade-off (Δsilhouette −0.093 ± 0.030 with Δ transfer up to +0.362 ± 0.096; Table 5). **(5) Domain.** The model is developed and validated on single-cell-resolution infectious-disease tissue — a regime of few, small, intensely heterogeneous slides — where the two-slide LODO and missing-modality stress tests are the realistic failure modes, rather than the many-sample tumour settings on which niche models are usually demonstrated. **(6) Translation.** The niche atlas is closed with a signature-reversal repurposing screen and an in-silico docking step against a host target, giving a testable host-directed hypothesis traceable back to specific niches (N4/N6 vs N0/N2) and markers. Read together, these six items define the paper's contribution: not a new attention operator, but a representation, a wiring, a training regime, a benchmarking discipline and a translational read-out assembled and stress-tested for a tissue setting in which none of them had been validated.

Relation to prior work. The closest prior art is **SpaCEy** (14), which established that learned niche representations are a productive abstraction in spatial transcriptomics. MuSpec differs in three ways that we consider substantive for the infectious-disease setting: it is applied to **granuloma**, not tumour, tissue; it is **unsupervised** and does not depend on annotated niche structure; and it is validated at **matched numbers of domains against a strong geometry-based baseline** with seed-controlled ablation and sensitivity, rather than at a self-selected operating point. NicheCompass (11) and Novae (12) are the nearest representation-learning and foundation-model neighbours; both are powerful and general, whereas MuSpec is deliberately small (0.78 M parameters) so that a 34-configuration sensitivity study is affordable on one consumer GPU. STAGATE (9), SpaGCN (8), and BANKSY (10) are the strong domain-detection baselines of the field; our matched-K result shows that MuSpec's advantage lies in **embedding separation**, and the moderate ARI/NMI agreement with BANKSY domains (0.06–0.23) shows that MuSpec adds information rather than re-deriving the geometric segmentation. GITIII (13) addresses cross-sample/cross-modal integration at a scale beyond this study; our two-slide LODO protocol is a small-scale but honest version of the same failure mode, and it is the reason we insist on joint training and report LODO ARI explicitly. Finally, scPerturBench (15) is the reason we adopted

matched-K and three-seed protocols at all: fair benchmarking is a design choice, and an unfair benchmark would have let us claim a larger effect.

Why the separation↔robustness trade-off matters. Two independent experiments — the ExternalAttention ablation and the 34-configuration hyperparameter sweep — converge on the same empirical structure: niche separation and robustness/transfer trade off against each other along a frontier, and the default configuration sits at the robust end. We regard this as the most useful methodological finding in the paper, precisely because it is the kind of result that a single-seed, single-configuration study would have hidden (it would simply have reported the high-separation configuration as "better"). Removing ExternalAttention lowered separation but raised robustness and transfer, which suggests the module sharpens the embedding at the cost of some invariance; the same trade-off appears as the reason that low-learning-rate and high-smoothing configurations, which maximise silhouette, degrade P1 robustness. Reporting the frontier rather than a single best number is what makes MuSpec useful as a component: a downstream user can choose a point on the frontier to match their tolerance for false niche boundaries.

Biological interpretation. The niches recovered on P1 are the most interpretable, and they read out a coherent granuloma architecture: an $SPP1^{+}$/$CTSK^{+}$ myeloid granuloma-core niche (N1), a fibrotic myofibroblast ring (N6), a retained airway/club-cell niche (N5), and an IL-11 fibroblast niche (N7). The cross-slide retention of the N5 airway identity and of a shared myeloid–fibroblast axis indicates that joint training captures conserved tissue structure, while the slide-specific markers on B1 reflect the greater internal heterogeneity of granuloma tissue. The SP110–IFN hub analysis connects the descriptive niche map to a known host TB-susceptibility protein, and the granuloma signature derived from the myeloid/fibrotic niches (N4/N6 vs N0/N2) is exactly the contrast one would expect to be therapeutically actionable.

Translational hypothesis. The signature-reversal screen nominates host-directed repurposing candidates, led by a **PPARγ-agonist triad** (pioglitazone, troglitazone, rosiglitazone), which is the most mechanistically coherent class for host-directed TB therapy: PPARγ signalling dampens excessive inflammation and supports macrophage antimicrobial function. The independent cross-checks (rituximab's rank-2 hit agreeing with the B-cell marker MS4A1 in B1 N7; the anti-fibrotic nintedanib docking strongly against SP110 and matching the N6 fibrotic niche) make the hypothesis more than a rank list.

Limitations

This study rests on two Stereo-seq slides — one granuloma and its peri-lesional margin — and the niche atlas describes these two specimens; the generality of the method itself is what the matched-K benchmark, ablation and sensitivity studies address. The 7-dimensional composition block is a deconvolution estimate (cell2location) cross-checked against measured spatial markers rather than ground-truth labels. The repurposing and docking outputs are computational and are offered as candidates for experimental follow-up.

Conclusions

MuSpec shows that a deliberately small, unsupervised multi-scale niche encoder — 782,251 parameters, trainable in minutes on a single consumer GPU — can recover biologically coherent and cross-slide-conserved spatial niches from single-cell-resolution Stereo-seq slides of human TB lung, and that its behaviour under input noise, missing modalities and cross-slide transfer can be benchmarked honestly rather than assumed. Applied to a pulmonary granuloma and its peri-lesional margin, it produced a nine-niche atlas with myeloid/granuloma, airway, fibrotic and fibroblast identities; an SP110–IFN hub hypothesis at the myeloid/granuloma end of the tissue; a 324-gene granuloma signature whose reversal nominates a PPARγ-agonist set for host-directed therapy; and in-silico docking against SP110 that ranks withaferin A and nintedanib as the strongest binders. The separation↔robustness frontier we report is a reminder that niche models should state where they are robust and where they are not, so that downstream users can choose an operating point that matches their tolerance for error. MuSpec's niche atlas and its nominated repurposing candidates are directly testable, including in high-burden TB settings.

List of abbreviations

ARI: adjusted Rand index; ASFF: adaptive spatial feature fusion; CP10K: counts per 10,000; ECM: extracellular matrix; FGN: Fourier graph network; GO-BP: Gene Ontology biological process; HVG: highly variable gene; IFN: interferon; kNN: k-nearest neighbour; LODO: leave-one-dataset-out; MSE: mean squared error; NMI: normalised mutual information; PC: principal component; scRNA-seq: single-cell RNA sequencing; TB: tuberculosis; UPR: unfolded protein response.

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Medical Ethics Committee of the Children's Hospital of Xinjiang Uygur Autonomous Region (Beijing Children's Hospital Xinjiang Hospital), approval number **KY2025022719**, and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Availability of data and materials

The two Stereo-seq slides analysed in this study are identified by BGI sample IDs (B1 = Y40012NC13; P1 = Y40012LA11). The raw spatial transcriptomic data are not deposited in a public repository in the current analysis frame and are available from the corresponding author on reasonable request. The derived analysis outputs (niche embeddings, marker lists, signature gene lists, repurposing scores and docking results) and the analysis code are available from the corresponding author on reasonable request. The peripheral-blood scRNA-seq reference cohort is used under the source terms governing that dataset; in the current analysis frame the reference carries no raw counts.

Software. Project name: MuSpec. Project home page: available from the corresponding author on reasonable request. Archived version: not applicable. Operating system(s): Linux (platform independent). Programming language: Python. Other requirements: PyTorch with CUDA. License: not applicable. Any restrictions to use by non-academics: none.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by (i) the Key R&D Program of Xinjiang, grant number **2025B03020-2**; (ii) the Xinjiang Second Medical College research programme, grant number **XGGK2025-01** (Principal Investigator: Prof. Jiangdong Wu, 2025.09–2028.09); and (iii) the Xinjiang Second Medical College research programme, grant number **ZR202656** (Principal Investigator: Tingting Luo).

Authors' contributions

J.W. conceived and supervised the study and secured funding. Y.G., T.L. and C.X. designed and performed the computational analyses. Y.G. assembled and trained the model, implemented the benchmarking, ablation and hyperparameter-sensitivity experiments, and drafted the manuscript. T.L. and C.X. performed the niche characterisation, differential-expression/signature construction and multi-library drug-repurposing analysis. J.C., Y.Z. and Y.Y. contributed data curation, cell-type annotation and figure preparation. Q.L. and Y.P. provided the clinical samples and contributed clinical and pathological interpretation. All authors read and approved the final manuscript.

Acknowledgements

We thank the authors of the ASFF, TripletAttention, ExternalAttention and Fourier-graph-network papers, whose published designs informed the four functional modules in this work. We thank the staff of the participating clinical centres for their assistance with sample collection and processing, and BGI for the Stereo-seq data generation.

References

References are formatted in Vancouver style, numbered in the order of first citation. Journal names are abbreviated per Index Medicus/MEDLINE.

1. World Health Organization. Global tuberculosis report 2025. Geneva: WHO; 2025. Licence: CC BY-NC-SA 3.0 IGO.
2. Ramakrishnan L. Revisiting the role of the granuloma in tuberculosis. *Nat Rev Immunol*. 2012;12(5):352–366. doi:10.1038/nri3211.
3. Ehlers S, Schaible UE. The granuloma in tuberculosis: dynamics of a host–pathogen collusion. *Front Immunol*. 2012;3:411. doi:10.3389/fimmu.2012.00411.
4. Marakalala MJ, Raju RM, Sharma K, et al. Inflammatory signaling in human tuberculosis granulomas is spatially organized. *Nat Med*. 2016;22(5):531–538. doi:10.1038/nm.4073.
5. McCaffrey EF, Donato M, Keren L, et al. The immunoregulatory landscape of human tuberculosis granulomas. *Nat Immunol*. 2022;23(2):318–329. doi:10.1038/s41590-021-01121-x.

6. Zumla A, Rao M, Wallis RS, et al. Host-directed therapies for infectious diseases: current status, recent progress, and future prospects. *Lancet Infect Dis*. 2016;16(4):e47–e63. doi:10.1016/S1473-3099(16)00078-5.

7. Chen A, Liao S, Cheng M, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays (Stereo-seq). *Cell*. 2022;185(10):1777–1792.e21. doi:10.1016/j.cell.2022.04.003.

8. Hu J, Li X, Coleman K, et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods*. 2021;18(11):1342–1351. doi:10.1038/s41592-021-01255-8.

9. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder (STAGATE). *Nat Commun*. 2022;13:1739. doi:10.1038/s41467-022-29439-6.

10. Singhal V, Chou N, Lee J, et al. BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nat Genet*. 2024;56(3):431–441. doi:10.1038/s41588-024-01664-3.

11. Birk S, Bonafonte-Pardàs I, Feriz AM, et al. Quantitative characterization of cell niches in spatially resolved omics data. *Nat Genet*. 2025;57(4):897–909. doi:10.1038/s41588-025-02120-6.

12. Blampey Q, Benkirane H, Bantsimba-Malanda C, et al. Novae: a graph-based foundation model for spatial transcriptomics data. *Nat Methods*. 2025;22(12):2539–2550. doi:10.1038/s41592-025-02899-6.

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. *Nat Mach Intell*. 2026;8(1):42–58. doi:10.1038/s42256-025-01161-0.

14. Rifaioglu AS, Ervin EH, Sarigun A, et al. SpaCEy links spatial tissue patterns to clinical outcomes using explainable graph neural networks. *Nat Commun*. 2026;17:10232. doi:10.1038/s41467-026-77924-z.

15. Wei Z, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nat Methods*. 2026;23(2):451–464. doi:10.1038/s41592-025-02980-0.

16. Kleshchevnikov V, Shmatko A, Dann E, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol*. 2022;40(5):661–671. doi:10.1038/s41587-021-01139-4.

17. Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection (ASFF). *arXiv*:1911.09516. 2019.

18. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. *Proc IEEE/CVF WACV*. 2021:3138–3147. doi:10.1109/WACV48630.2021.00318.

19. Guo MH, Liu ZN, Mu TJ, Hu SM. Beyond self-attention: external attention using two linear layers for visual tasks. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(5):5436–5447. doi:10.1109/TPAMI.2022.3211006.

20. Yi K, Zhang Q, Fan W, et al. FourierGNN: rethinking multivariate time series forecasting from a pure graph perspective (FGN). *Adv Neural Inf Process Syst (NeurIPS)*. 2023.

21. Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell–cell communication using CellChat. *Nat Commun*. 2021;12:1088. doi:10.1038/s41467-021-21246-9.

22. Pan H, Yan BS, Rojas M, et al. Ipr1 gene mediates innate immunity to tuberculosis (SP110). *Nature*. 2005;434(7034):767–772. doi:10.1038/nature03419.

23. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171(6):1437–1452.e17. doi:10.1016/j.cell.2017.10.049.

24. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313(5795):1929–1935. doi:10.1126/science.1132939.

25. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J Chem Inf Model*. 2021;61(8):3891–3898. doi:10.1021/acs.jcim.1c00203.

26. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455–461. doi:10.1002/jcc.21334.

27. Almeida PE, Carneiro AB, Silva AR, Bozza PT. PPARγ expression and function in mycobacterial infection: roles in lipid metabolism, immunity, and bacterial killing. *PPAR Res*. 2012;2012:383829. doi:10.1155/2012/383829.

28. Schafer S, Viswanathan S, Widjaja AA, et al. IL-11 is a crucial determinant of cardiovascular fibrosis. *Nature*. 2017;552(7683):110–115. doi:10.1038/nature24676.

29. Qi J, Sun H, Zhang Y, et al. Single-cell and spatial analysis reveal interaction of FAP⁺ fibroblasts and SPP1⁺ macrophages in colorectal cancer. *Nat Commun*. 2022;13:1742. doi:10.1038/s41467-022-29366-6.

30. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.

Note. Authors are asked to confirm each entry against the publisher record before submission.

Figure legends

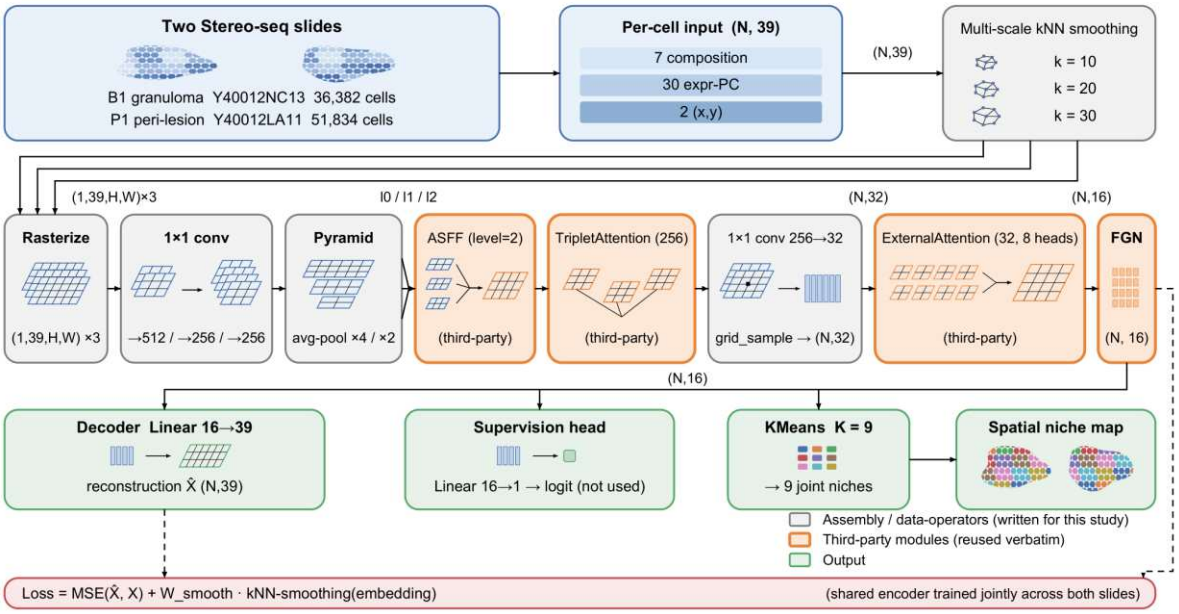


Figure 1. MuSpec architecture and computational cost. Schematic of the MuSpec forward pass. A 39-dimensional per-cell input (7-dimensional cell-type composition ‖ 30 expression PCs ‖ 2 spatial coordinates) is smoothed by multi-scale kNN averaging at three neighbourhood scales (k = 10/20/30) and rasterised onto a ≤224² lattice to give three (1, 39, H, W) tensors; 1×1 convolutions project each scale, an average-pooling pyramid aligns them, and ASFF(level = 2) performs adaptive cross-scale fusion. TripletAttention(256) applies spatial attention; a 1×1 convolution compresses the fused map to 32 channels, from which each cell's feature is

sampled with grid_sample at its fractional coordinate. ExternalAttention(32, 8 heads) then applies external-memory channel attention, and the Fourier graph network (FGN) produces the 16-dimensional niche embedding, which a linear decoder maps back to the 39-dimensional reconstruction ($\hat{X}$). KMeans (K = 9) on the embedding yields the nine joint niches and the spatial niche map. Box colours distinguish the pipeline code written for this study, the four functional modules that implement published attention-and-fusion architectures (Table 2b), and the outputs. Training minimises MSE($\hat{X}$, X) + W_smooth · kNN-smoothing(embedding) with a shared encoder trained jointly across both slides. The schematic also depicts a linear read-out head (16 → 1 → logit) that is not used in the unsupervised losses of this study (see Methods, MuSpec architecture). The model comprises 782,251 trainable parameters and is trained for 500 steps (the selected default budget; see Methods) on a single consumer GPU (≈7 GB peak; Table 7).

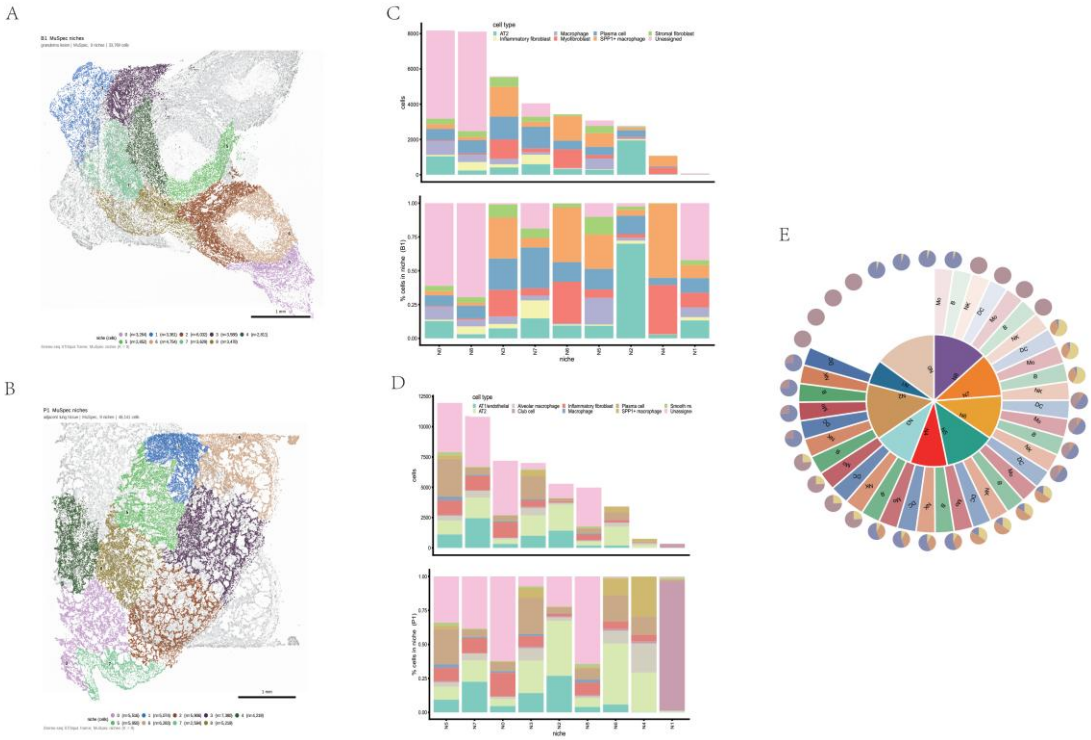


Figure 2. MuSpec recovers conserved spatial niches on a granuloma and its peri-lesional margin. (a) Nine MuSpec niches (K = 9) on the granuloma slide B1, shown in situ (33,769 cells). (b) The same nine joint niches on the peri-lesional slide P1 (48,141 cells). (c) Per-niche cell-type composition on B1, from the spatial cell-type annotations. (d) Per-niche cell-type composition on P1. (e) Sunburst summary of niche × blood-reference-class composition (DC, dendritic cells; NK, NK cells; B, B cells; Mo, monocytes). Niches are shared across slides by construction of the jointly trained encoder.

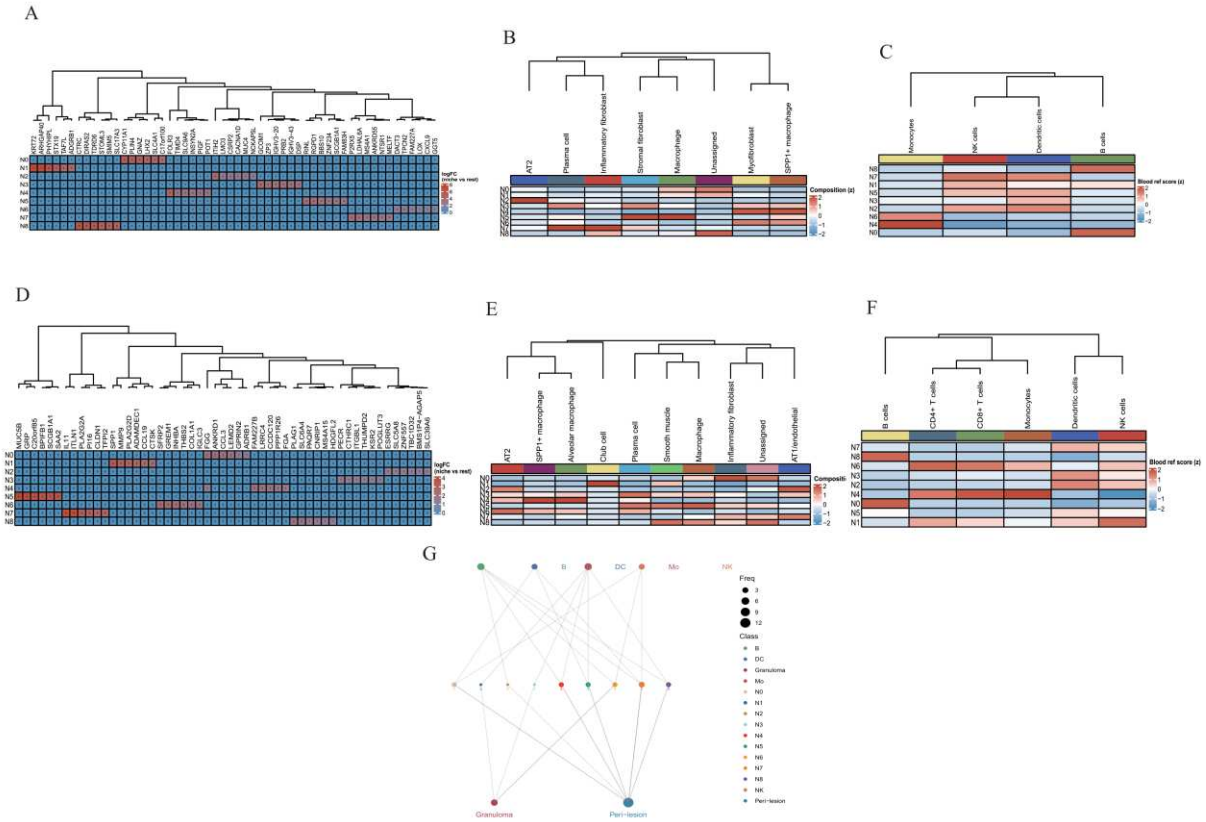


Figure 3. Biological characterisation of the joint niches. (a) Top-six marker genes per niche on B1 (log2FC versus the rest; technical genes removed; candidates detected in ≥10% of cells). (b) Niche × cell-type composition heatmap (z-scored) on B1. (c) Niche × blood-reference-class score heatmap on B1. (d) Top-six marker genes per niche on P1. (e) Niche × cell-type composition heatmap on P1. (f) Niche × blood-reference-class score heatmap on P1. (g) Niche–cell-type association network, with nodes coloured by granuloma versus peri-lesion membership.

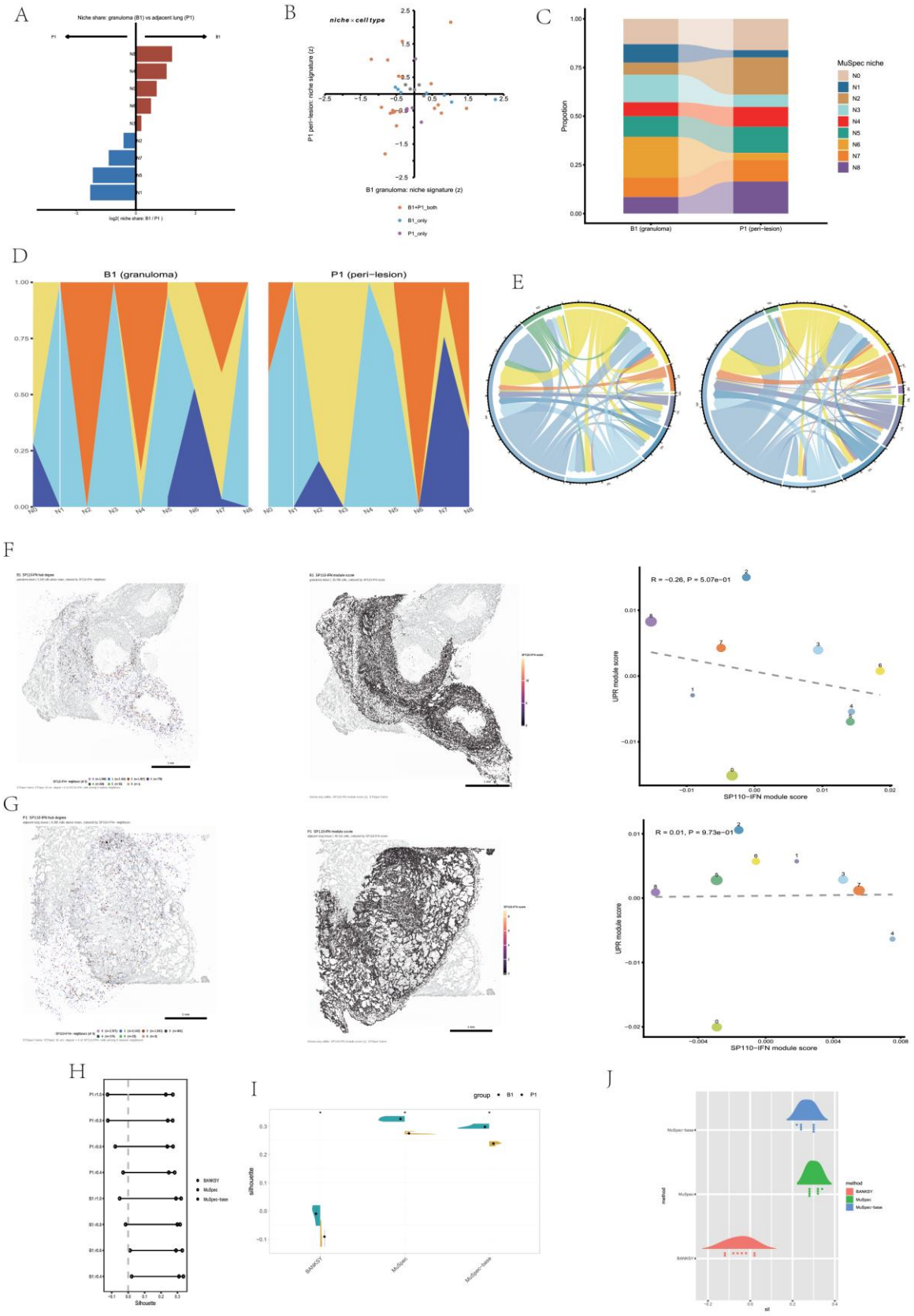


Figure 4. Cross-slide conservation, niche communication, and the SP110–IFN hub, with a matched-K separation benchmark. (a) Niche share on the granuloma (B1) versus peri-lesional (P1) slide, as log2(B1/P1) share per niche. (b) Niche-signature correlation between B1 and P1 (z-scored niche signatures), coloured by

whether a signature is shared (B1+P1_both) or slide-specific (B1_only, P1_only). (c) Alluvial (ribbon) diagram linking each niche's cell-type composition proportion on B1 to the same niche on P1, so that the width of each ribbon is the niche composition proportion and the change in ribbon width between the two columns visualises reallocation (the proportion denominator is the total cells of the slide, and niches are linked one-to-one by shared niche identity). (d) Per-niche composition proportions as stacked areas, faceted by slide (B1, P1). (e) Niche-to-niche ligand–receptor chord diagrams for B1 and P1. (f) SP110–IFN hub degree (number of SP110–IFN^{+} cells among the six spatial nearest neighbours) mapped for B1 and P1. (g) SP110–IFN module score mapped for B1 and P1, with insets showing the relationship between the SP110–IFN and UPR module scores (B1: R = 0.01, P = 9.73×10^{-1}; P1: R = −0.26, P = 5.07×10^{-1}). (h) Silhouette across the eight matched-K conditions (r0.4–r1.0 on B1 and P1) for BANKSY, MuSpec, and the single-scale base (MuSpec-base), with significance marks. (i) Box/violin summary of the same silhouette comparison, grouped by slide. (j) Density of silhouette by method.

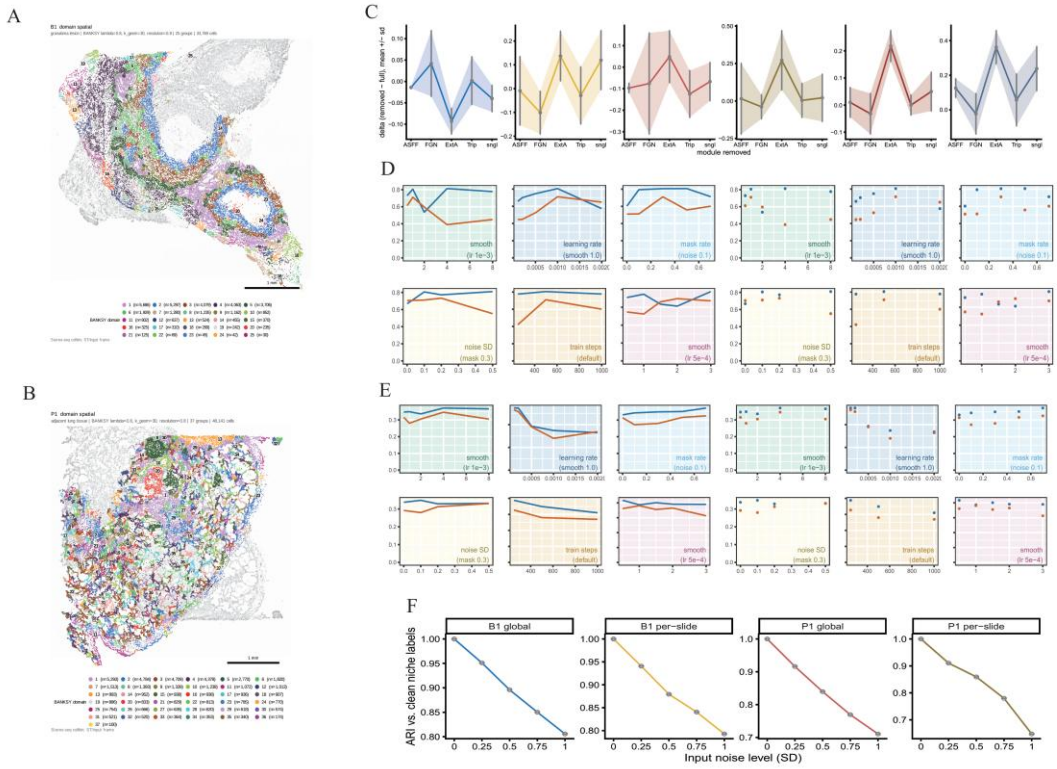
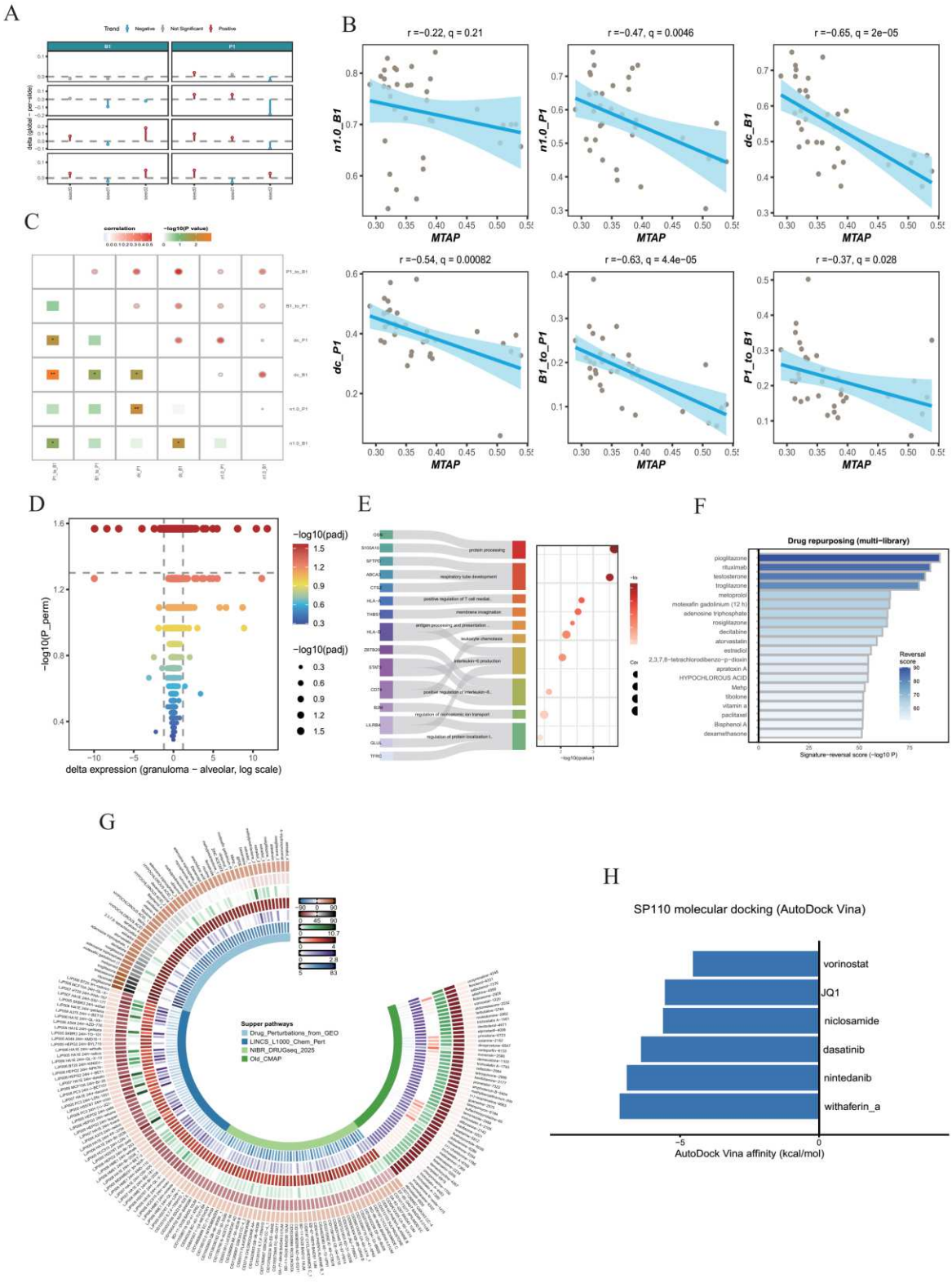


Figure 5. Matched-K domain segmentation, module ablation, and hyperparameter sensitivity. (a) BANKSY spatial domains on the granuloma slide B1 (λ = 0.8, k_geom = 30, resolution = 0.8; 25 domains; 33,769 cells), for qualitative comparison with the K = 9 MuSpec niches of Fig. 2a. (b) BANKSY spatial domains on P1 (37 domains; 48,141 cells). (c) Three-seed module ablation: Δ (module removed − full model),

mean ± s.d., for six metrics (silhouette; noise-robustness ARI; drop-composition ARI on B1 and P1; B1→P1 and P1→B1 transfer ARI) across the five ablations (ASFF, FGN, ExternalAttention, TripletAttention, single-scale). (d, e) Single-factor sensitivity of silhouette to smoothing, learning rate, mask rate, noise SD, training steps, and a smooth × learning-rate knee configuration (d: default grid; e: grid with the additional smooth/lr 5e-4 row). (f) Noise-robustness (ARI versus the clean niche labels) as a function of input-noise SD, for global versus per-slide normalisation on B1 and P1.



871

872 **Figure 6. Robustness summary and translation: granuloma signature, host-directed repurposing, and**

873 **docking.** (a) Paired Δ(global − per-slide) normalisation for four metrics (silhouette; noise-robustness ARI;

874 drop-composition ARI; LODO transfer ARI) on B1 and P1 across seeds 0/1/2, showing sign flips for most

metrics. (b) Scatter of each configuration's mean silhouette separation $MTAP = \bar{s} = ½(s_B1 + s_P1)$ against each of the six robustness metrics (Spearman r and q annotated). (c) Correlation bubble heatmap (−log10 P) summarising the same relationships. (d) Volcano plot of the 324-gene granuloma signature (granuloma minus alveolar; permutation and BH-adjusted P). (e) GO biological-process enrichment of the signature (−log10 q; dot size = gene count). (f) Top signature-reversal hits from the multi-library repurposing screen (GEO perturbations, LINCS L1000 chemical perturbations, NIBR DRUGseq 2025, and old CMAP, merged). (g) Circos linking repurposing hits to their source libraries (LINCS L1000 chemical perturbations, GEO perturbations, NIBR DRUGseq 2025, old CMAP) and super-pathways. (h) SP110 molecular docking of the focused six-ligand panel (AutoDock Vina affinity, kcal/mol).

Supplementary Figures. S1. Niche–cell-type network (alternative presentation of Fig. 3g). S2. Weighted ligand–receptor chord diagrams, B1 and P1 (alternative presentation of Fig. 4e). S3. Ligand–receptor bubble plots, B1 and P1. S4. Alternative silhouette presentation for the three methods (companion to Fig. 4h, i). S5. Single-panel sensitivity curves (silhouette; noise-robustness) for the hyperparameter sweep (companion to Fig. 5d, e). S6. Robustness curves versus perturbation magnitude (companion to Fig. 5f). S7. Cell-type × niche characteristic overview. S8. Alternative matched-K presentation if the main figure uses the swarm layout.