

Conversational analysis of spatial data: bypassing the biostats bottleneck with foundation models and LLMs.



Scan here to download or learn more

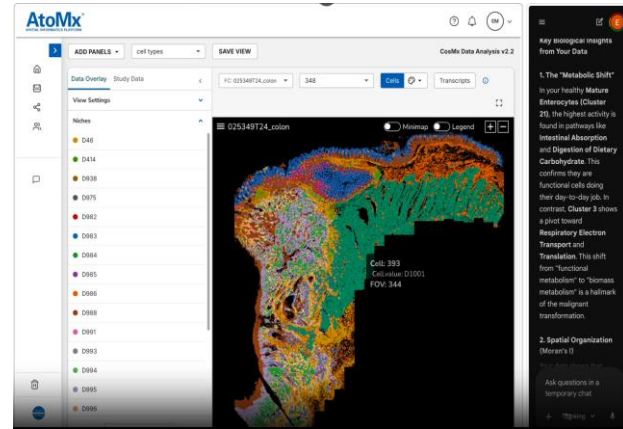


Patrick Danaher, Evelyn Metzger, Nicole Ortogero, Stacey Walker, Sayani Bhattacharjee, Prajan Divakar, Mirko Corselli, Joseph M. Beechem

Bruker Spatial Biology

The CosMx-AtoMx[®] Spatial Informatics Platform

- CosMx multi-omic whole transcriptome imaging data is so complete (informationally), that when the high-dimensional data is "packaged-properly" for LLMs (work that we accomplished) that exceptional first-pass conversation-driven analysis can be performed directly by the biological-expert (not the informaticians) Our goal: empower biologists to perform high-plex spatial analyses, quickly and simply



How can AI speed data analysis?

1. Large language models can help explore data and place results in context.
2. Deep learning can perform the difficult tasks of cell typing and neighborhood annotation.

What about analysis agents? Analysis agents aren't reliable enough (as of Feb 2026) to trust with fully-automated analyses. Human-LLM collaborations are more reliable and conducive to scientific insight.

Setting up LLMs to succeed as collaborators



AtoMx automatically processes your data, then creates a data package full of useful summary statistics optimized for ingestion by LLMs.

Ask the LLM for interpretive help:

- What are these clusters / niches?
- Which pathways look interesting?
- What morphological features are present in this tissue?

Validate in the AtoMx viewer, Form new questions for the LLM

Deep learning on the AtoMx database adds value to every new study

Proposition: 0-shot embeddings can align cell types or cellular neighborhoods across tissue types and batches. The nearly 3 billion cells on AtoMx are the world's best training database for such embeddings.

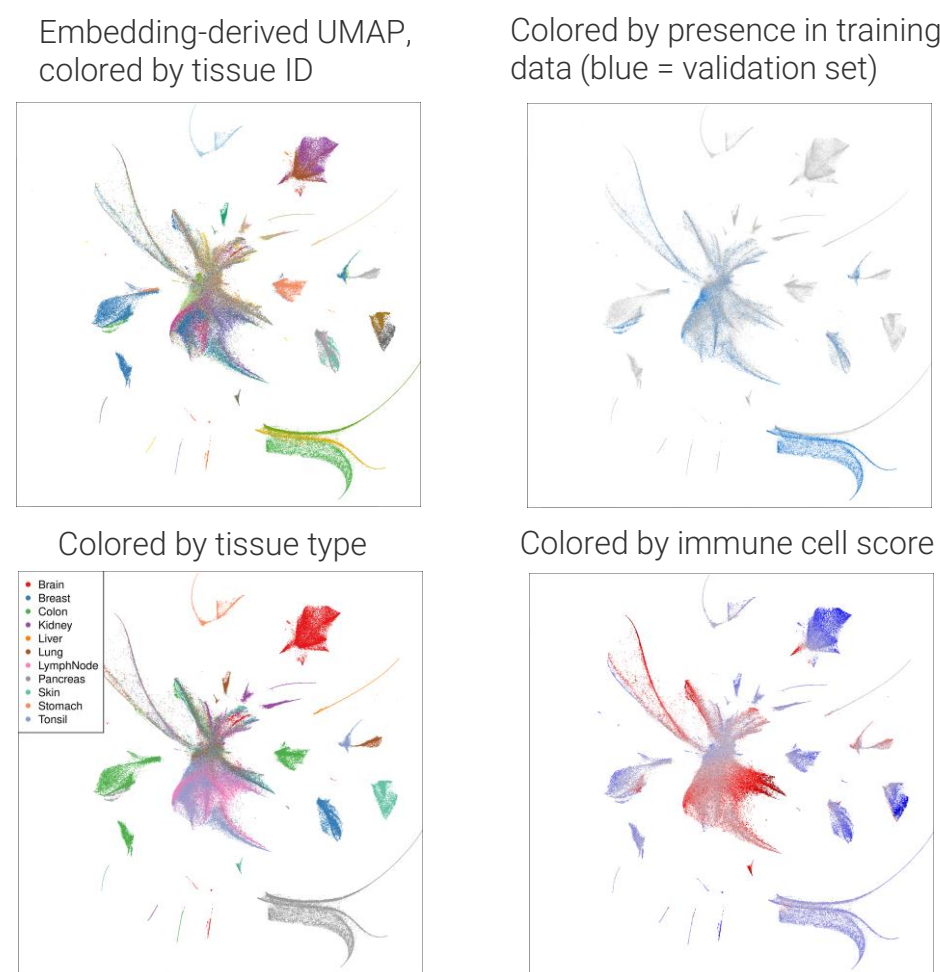
We are developing **automatic annotation engines** that classify cells / neighborhoods based on their 0-shot embeddings.

Our goal: every new study should benefit from the data on AtoMx.

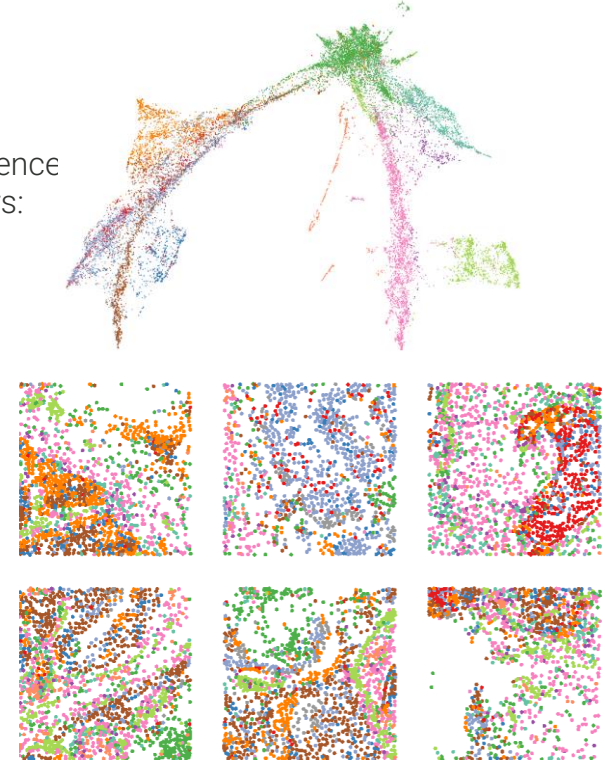
Proof of concept: annotating cells by nearest neighbors in 0-shot embedding

We used 23 tissues (860,000 cells) to train a 50-dimension cell embedding.

We then applied this embedding to a breast tumor not used in training. We annotated cells based on their nearest neighbors in the 2 breast cancer datasets used in training.



UMAP on 0-shot embedding, with cell types called by reference set nearest neighbors:



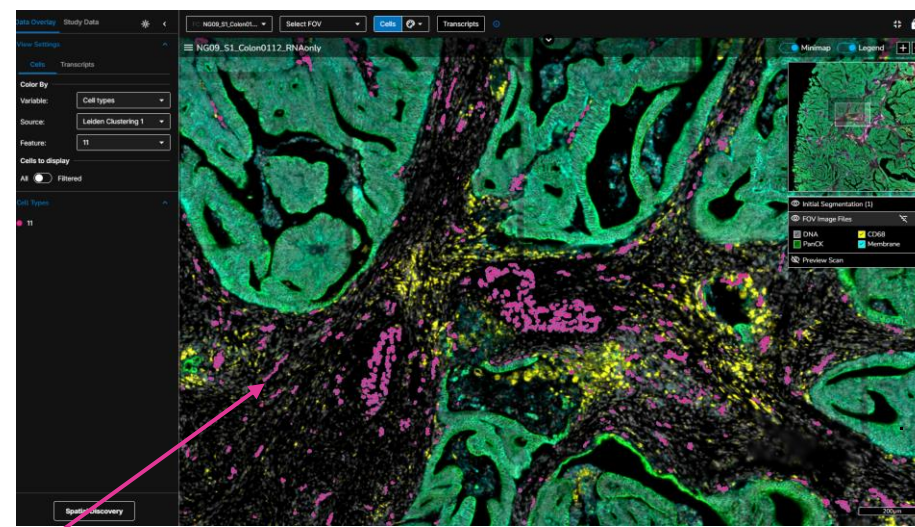
Spatial arrangement of reference-derived cell types:

Note: the above results are 0-shot. There was no batch correction and no learning from the new dataset.

Example: LLM-assisted cell type annotation

Please try to identify the cell type behind each cluster. Present your results as a table, returning: the cluster name, the proposed cell type, the relevant marker genes behind your decision, your confidence, and the nature of the expected spatial distribution

Cluster	Proposed cell type	Marker genes driving call (examples)	Confidence	Expected spatial distribution
1	Tumor epithelial (WNT/EGFR-ish program)	AXIN2, RNF43, AREG, OLFML4, DPEP1, ANO9	Medium	Tumor nests; likely patchy subregions within epithelium
2	Antigen-presenting myeloid (macrophage/DC-like)	LAPTM5, CD74, HLA-DRA, HLA-DRB1, HLA-DQB1, CCL4, APOC1	High	Stromal immune areas; often enriched near lymphoid aggregates / TLS-like regions
3	T cells (CXCR4+, activated/chemokine-associated)	TRBC1, PTPRC, LTB, CORO1A, LCP1, CXCR4	High	Immune infiltrates; scattered or clustered in TLS/peritumoral rims
11	Vascular endothelial (possibly mixed with mural/perivascular signal)	PECAM1, VWF, PLVAF, EGFL7, COL4A1, COL4A2, SPARCL1 (plus NOTCH3/MCAM)	High	Along vessels; linear/branching structures; enriched near perfused stroma

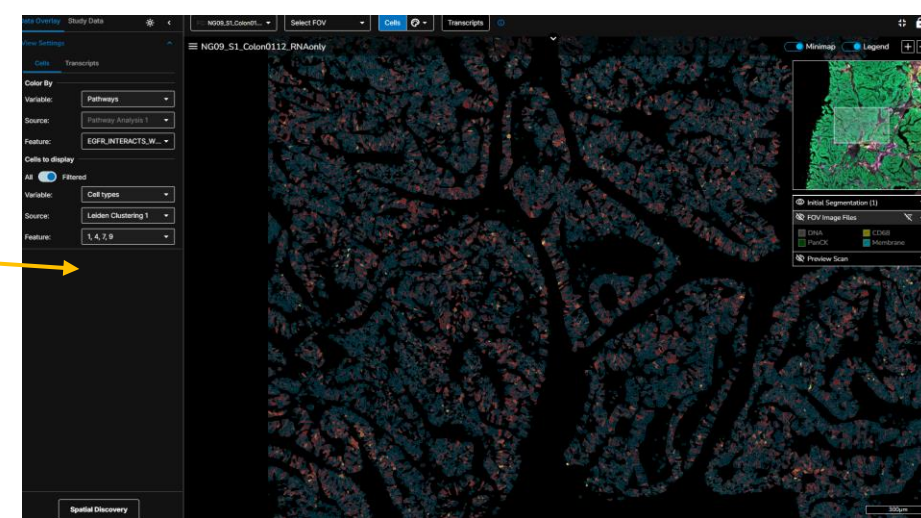


Example: LLMs suggest directions for exploratory analysis

Now look for pathways that are spatially variable in cluster 1, either that are enriched differently across niches, or that have high Moran's I. Suggest pathways that might be interesting to explore further, based on both 1. having spatial trends and 2. having some biological relevance to colon cancer.

Candidate spatially variable pathways in Cluster 1

Pathway (Reactome)	Why interesting biologically (colon cancer relevance)	Evidence of spatial structure	Likely spatial pattern
EGFR interacts with phospholipase C gamma (PLCγ)	EGFR signaling is a core driver of colorectal cancer (CRC); links to proliferation, invasion, and therapy response (anti-EGFR drugs)	High Moran's I, FDR-significant	Localized tumor subregions with strong growth factor signaling (likely tumor edge or niche-specific hotspots)
PI3K cascade (GFRβ)	PI3K/AKT pathway frequently activated in CRC; supports survival and metabolic adaptation	High Moran's I	Patchy activation across tumor nests, possibly linked to stromal contact or hypoxia
E2F-mediated regulation of DNA replication	E2F drives cell-cycle entry; hallmark of proliferative tumor compartments	Very high Moran's I (>1)	Proliferative foci rather than uniform tumor expression
Inhibition of replication initiation	Reflects spatial control of cell-cycle checkpoints / stress responses	Very high Moran's I	Spatially segregated cycling vs non-cycling tumor regions



Coming soon: mini data packages for custom exploration

Can you give me code to reproduce that analysis?

```
python
import pandas as pd
import numpy as np
from itertools import combinations
from scipy.stats import mannwhitneyu

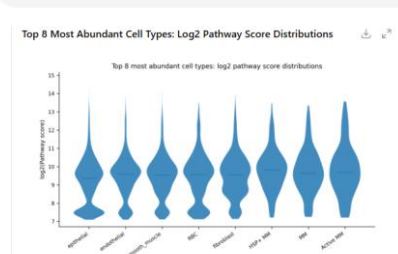
# ... load ...
cxk_path = "test/data/example_data_package.csv" # change as needed
mm_clusters = ["MM", "Active MM", "HSP-MM"] # must match your 'cluster' labels exactly

# ... load ...
df = pd.read_csv(cxk_path)

# basic checks
required_cols = ["cluster", "pathway_score"]
missing = required_cols - set(df.columns)
if missing:
    raise ValueError(f"Missing required columns: {missing}. Found columns: {list(df.columns)}")
```

Select cells and variables of interest, then ask LLMs to code up custom plots and analyses.

Can you make me a violin plot of log2 pathway score for the most abundant 8 cell types?



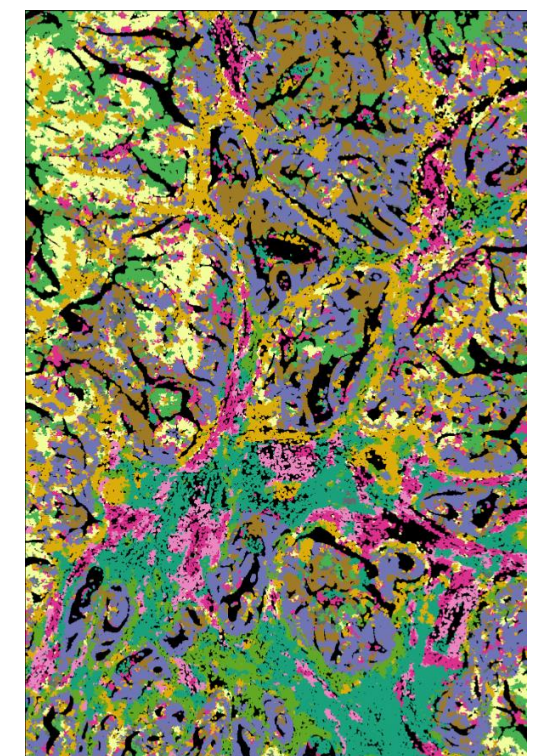
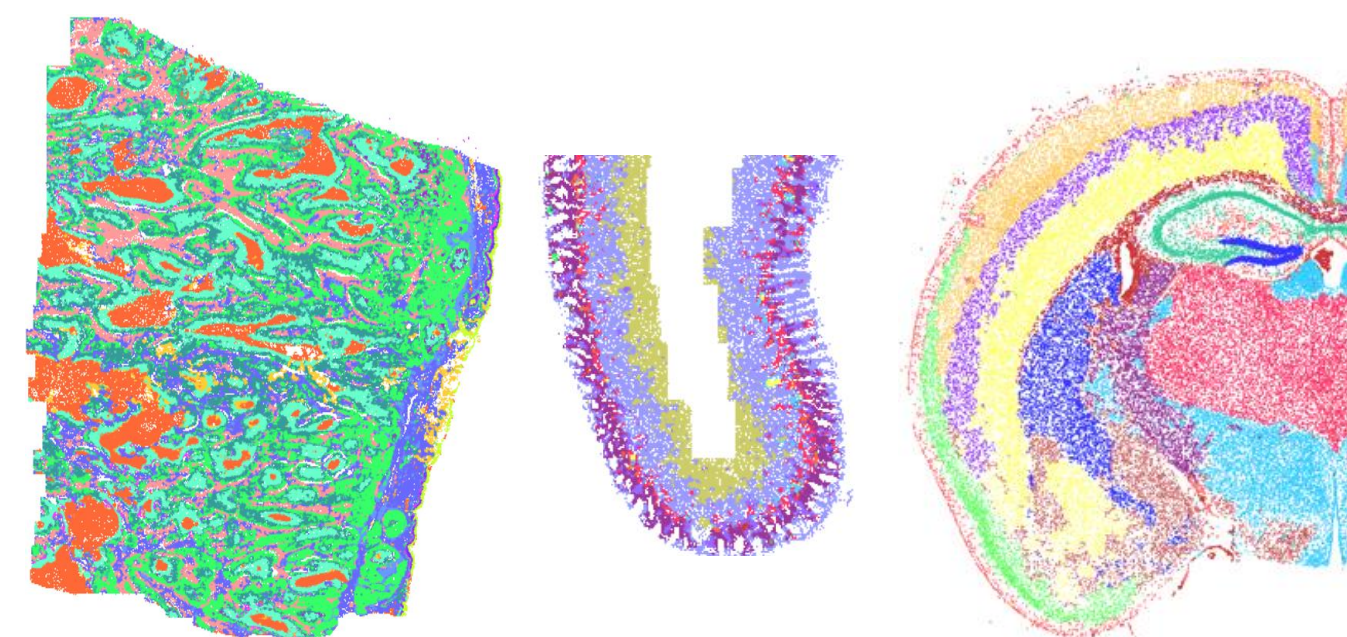
Can you give me some statistics on the differences between the MM clusters in the above plot?

Cluster	n	Mean	Median	SD	IQR (25-75%)
MM	447	9.72	9.64	1.24	8.95-10.66
Active MM	350	9.80	9.69	1.41	8.90-10.60
HSP-MM	459	9.81	9.81	1.16	9.04-10.43

Key point: all three distributions are very similar in location and spread.

CosMx-trained Novae model defines 0-shot niches in CosMx datasets

Novae is a foundation model for classifying spatial domains. We fine-tuned iNovae's foundation model using 98 CosMx studies. AtoMx uses this fine-tuned model to power a consistent niche classification schema across tissue types. (A separate brain model used 12 studies.)



Article | Published: 10 December 2025
Novae: a graph-based foundation model for spatial transcriptomics data
 Quentin Blumway ^{1,2}, Isakim Benkhanou, Nadège Benoit, Kevin Müller, Grégoire Geslain, Florent Gehlen, Fabrice André & Paul Henry Comte ^{1,2,3}
 Nature Methods | 22: 2539-2550 (2025) | [View this article](#)