The Journal of Open Source Software

# MOSAIK: Multi-Origin Spatial Transcriptomics Analysis and Integration Kit

**Anthony Baptista** [1,2], **Diane Cruiziat**[1], **Rosamond Nuamah** [1], **Ciro Chiappini** [3,4], **and Anita Grigoriadis** [1]

**1** Cancer Bioinformatics, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom **2** The Alan Turing Institute, The British Library, London, NW1 2DB, United Kingdom. **3** Centre for Cranio facial and Regenerative Biology, King's College London, London, SE1 9RT, United Kingdom **4** London Centre for Nanotechnology, King's College London, London WC2R 2LS, United Kingdom.

**DOI:** 10.21105/joss.08795

**Software**

- Review
- Repository
- Archive

**Editor:** William Gearty

**Reviewers:**

- @skanwal
- @haessar

**Submitted:** 16 May 2025
**Published:** 16 January 2026



## Summary

Spatial transcriptomics (ST) has revolutionised transcriptomics analysis by preserving tissue architecture, allowing researchers to study gene expression in its native spatial context. However, despite its potential, ST still faces significant technical challenges. Two major issues include: (1) the integration of raw data into coherent and reproducible analysis workflows, and (2) the accurate assignment of transcripts to individual cells. To address these challenges, we present MOSAIK, the first fully integrated, end-to-end workflow that supports raw data from both NanoString CosMx Spatial Molecular Imager (CosMx) and 10x Genomics Xenium In Situ (Xenium). MOSAIK (Multi-Origin Spatial Transcriptomics Analysis and Integration Kit) unifies transcriptomics and imaging data into a single Python object based on the spatialdata format. This unified structure ensures compatibility with a broad range of Python tools, enabling robust quality control and downstream analyses. With MOSAIK, users can perform advanced analyses such as re-segmentation (to more accurately assign transcripts to individual cells), cell typing, tissue domain identification, and cell-cell communication within a seamless and reproducible Python environment.

## Statement of need

Spatial transcriptomics (ST) enables the study of transcriptomes within intact tissues, which is essential for understanding a cell's position relative to its neighbours and the surrounding extracellular structures. This spatial context provides crucial insights into cellular phenotype, function, and disease progression, particularly in cancer, where the tumour micro-environment (TME) influences processes such as chemo-resistance (Mehraj et al., 2021). The commercialisation of ST platforms has expanded access to these technologies, earning ST the title of "Method of the Year 2020" by Nature Methods (Marx, 2021).

Imaging-based fluorescence in situ hybridisation (FISH) technologies provide high-multiplex, subcellular-resolution transcriptomics data across over one million cells. These platforms, such as CosMx by NanoString and Xenium by 10x Genomics, offer high sensitivity and specificity, facilitating the exploration of cell atlases, cell–cell interactions, and the phenotypic architecture of the TME (Chen et al., 2015; Vandereyken et al., 2023).

Despite its promise, ST faces two primary technical challenges: (1) integrating raw ST data into standardised, reproducible workflows—complicated by variability in platforms and data formats; and (2) accurately assigning transcripts to individual cells within heterogeneous tissue architectures. These challenges hinder downstream analyses including cell type identification,

spatial gene expression mapping, and cell-cell interaction inference. The diversity of ST technologies offers distinct advantages: some provide higher spatial resolution, others greater transcriptomic depth or tissue-type compatibility, making disparate individual platforms well-suited to specific biological questions. However, a unified workflow accommodating multiple platforms, from raw data processing to downstream analysis, is still needed to streamline cross-platform integration and enable effective multimodal analyses.

To address the first challenge, we developed a unified workflow that supports raw data from both CosMx and Xenium. While a Xenium reader already exists and handles multiple modalities effectively, CosMx readers lacked robustness in several areas: handling of coordinate systems, creation of segmentation polygons, and reintegration of multi-channel images. We addressed these limitations to ensure the resulting Python object matches the Xenium output format. We also aligned the workflow with the most suitable Python library, *SpatialData* (Marconato et al., 2025), which integrates spatial elements (images, transcript locations, cell segmentation labels and shapes (polygons)) with transcriptomics data into an annotated dataframe suitable for single-cell analysis.

Addressing the second challenge requires precise cell segmentation, which directly affects downstream analysis accuracy. Our workflow integrates native segmentation approaches: CosMx uses Cellpose (Stringer et al., 2021), while Xenium employs Voronoi expansion (Janesick et al., 2023). Users may also apply alternative or custom segmentation tools which can offer improved performance but typically require careful parameter tuning. Such tuning is difficult to implement in tools like Xenium Ranger (10x Genomics) or AtoMx (NanoString).

This integrated pipeline provides a foundation for downstream modelling and analysis, offering a scalable solution for tackling key challenges in ST, especially in multimodal data integration.

## Overview of the workflow

The MOSAIK workflow (https://github.com/anthbapt/MOSAIK) supports both CosMx and Xenium ST platforms through modular pipelines designed for data integration, visualisation, and analysis (Fig. 1). For CosMx, data are first exported from the AtoMx platform, including all Flat Files and relevant raw files such as Morphology2D. As a first step, the data may be visualised in Napari according to the guidelines outlined in the documentation. As a second step, the data exported from AtoMx are uncompressed and organised using helper scripts to generate structured directories (e.g., CellComposite, CellLabels) essential for downstream processing.

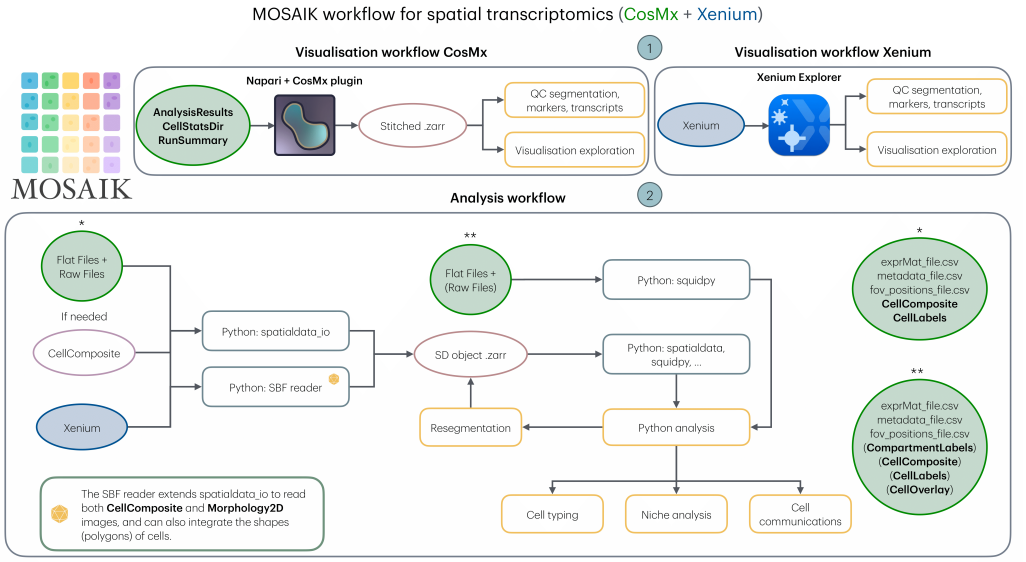


**Figure 1:** The MOSAIK workflow is divided into two parts: the visualisation component, which enables quality assessment of the immunofluorescence staining and verification of cell segmentation; and the data integration component, which leads to downstream analysis. 1: The MOSAIK visualisation is based on two visualisation strategies: On one hand, Napari with the CosMx plugin to visualise CosMx data; on the other hand, Xenium Explorer for Xenium data. 2: MOSAIK analysis takes the raw data and converts it into a Python object, making it easy to perform quality control and facilitate downstream analysis.

Structured inputs are then read into the analysis pipeline using a custom reader, which extends the spatialdata_io framework to incorporate various image types along with cell shape annotations (polygons). These information are stored into a Zarr file, which is an open standard for storing large multidimensional array data. Then, the resulting Zarr object is processed using Python-based tools such as squidpy and spatialdata for quality control and downstream analyses, including re-segmentation, cell typing, niche identification, or cell-cell communication.

Xenium data follow a similar pipeline. The data can be visualised using the software provided by 10x Genomics, Xenium Explorer. The data exported directly from the instrument can then be processed through the reader embedded in MOSAIK and converted into a Zarr file. This unified format is subsequently analysed using the same set of Python tools, ensuring consistency across platforms.

MOSAIK is the first fully integrated end-to-end workflow that supports both CosMx and Xenium raw data, standardising their output into a unified spatial data format (Fig. 2). The entire process is thoroughly documented in the MOSAIK GitHub repository, which includes two example workflows: one using a publicly available CosMx dataset from the NanoString website, and another using a Xenium dataset from the 10x Genomics platform.

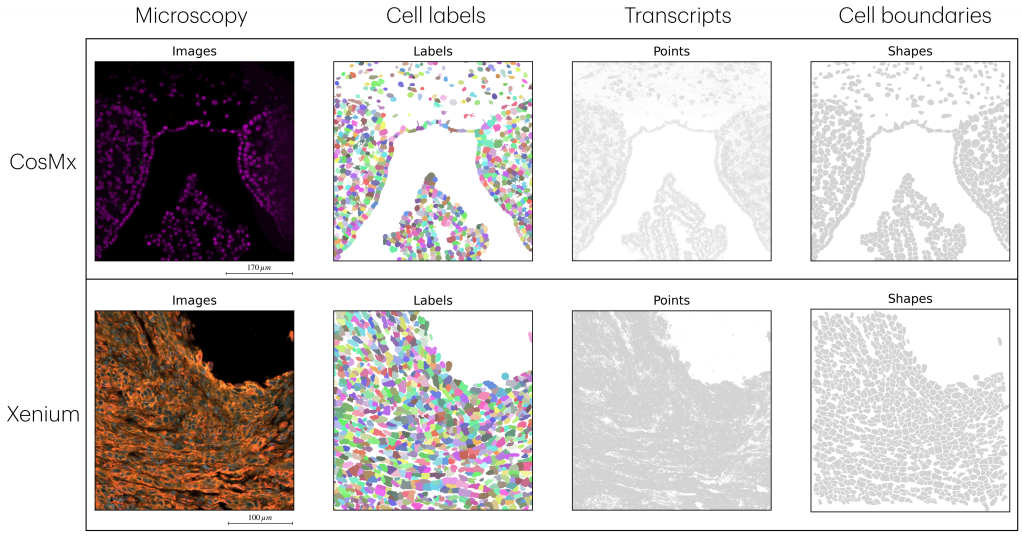


**Figure 2:** The Python SpatialData object obtained after using the MOSAIK workflow embeds both CosMx and Xenium data into similar objects, which can be combined or compared. MOSAIK, along with the Python library *SpatialData*, allows for the visualisation and connection of SpatialElements: Images (e.g., H&E or immunofluorescence stains), Labels (segmentation maps), Points (i.e., transcripts), and Shapes (e.g., cell/nucleus boundaries or ROIs). The first two objects are raster objects (images), and the last two are vector objects (points and polygons). The CosMx fields of view are defined by a $510\,\mu\mathrm{m}$ square box, and for Xenium, each pixel represents $0.2125\,\mu\mathrm{m}$. Both CosMx and Xenium data are sourced from public repositories (see the Data availability section).

Finally, we have created a GitHub repository (https://github.com/anthbapt/Spatial-Biology-Tools) that compiles a collection of Python tools designed to be used alongside or after integration with our workflow. These tools support a wide range of applications, including segmentation, cell typing, domain identification, gene imputation, detection of spatially variable genes, cell–cell communication analysis, dimensionality reduction, multimodal integration, and the use of foundation models, among others. By providing this curated collection, our goal is to guide users seamlessly from raw data to advanced analytical applications, all within a unified and community-supported framework.

## Data availaility

The datasets used to generate the figures are publicly available at the following websites:

- https://nanostring.com/cosmx-mouse-brain-ffpe
- https://www.10xgenomics.com/xenium-prime-ffpe-human-skin

The processed datasets associated with jupyter notebooks are provided as examples in the following Zenodo repository: https://doi.org/10.5281/zenodo.17700741.

## Code availaility

The MOSAIK workflow is publicly available on GitHub at https://github.com/anthbapt/MOSAIK.

## Related software

This work integrates well with the existing spatial transcriptomics (ST) community, particularly with tools that are part of the scverse ecosystem (Virshup et al., 2023), such as *scanpy* (Wolf et al., 2018) and *spatialdata* (Marconato et al., 2025), which provide the core Python objects used as the foundation of MOSAIK, namely *AnnData* and *SpatialData*.

Moreover, well-established libraries such as *squidpy* (Palla et al., 2022) and *scvi-tools* (Gayoso et al., 2022) can be easily incorporated into downstream analyses. In addition, we provide seamless integration of *Cellpose-SAM* (Pachitariu et al., 2025) and *Proseg* (Jones et al., 2025) for cell segmentation.

## Planned Enhancements

Recognising that ST is a rapidly evolving field, MOSAIK is designed to remain aligned with the latest standards, both in terms of experimental setup and raw data processing, as well as on the computational side by integrating emerging methods and tools in development. We are planning to allow multimodal integration; modalities currently under consideration include H&E, Akoya PhenoCycler, IMC, and metallomics data.

## Author Contributions

- Anthony Baptista: Developed the MOSAIK workflow and wrote the manuscript.
- Diane Cruiziat: Co-developed the segmentation part of the MOSAIK workflow.
- Rosamond Nuamah: Provided feedback on the spatial transcriptomics data and highlighted the experimental requirements to ensure the numerical workflow aligns with the raw data.
- Anita Grigoriadis and Ciro Chiappini: Lead the King's College London Spatial Biology Facility, where this work was conducted, and proofread the manuscript.

## Acknowledgements

Anthony Baptista, and Anita Grigoriadis acknowledge support from the CRUK City of London Centre Award [CTRQQR-2021/100004]. Anthony Baptista, Diane Cruiziat, Rosamond Nuamah, Ciro Chiappini, and Anita Grigoriadis acknowledge support from MRC [MR/X012476/1].

## References

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, *348*(6233), aaa6090. https://doi.org/10.1126/science.aaa6090

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., … Yosef, N. (2022). A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, *40*(2), 163–166. https://doi.org/10.1038/s41587-021-01206-w

Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Williams, S. R., Rouault, M., Beliakoff, G., Morrison, C. A., Oliveira, M. F., Sicherman, J. T., Kohlway, A., Abousoud, J., Drennon, T. Y., Mohabbat, S. H., Taylor, S. E. B., & Development Teams, 10x. (2023). High resolution mapping of the tumor microenvironment using integrated single-cell, spatial

and in situ analysis. *Nature Communications*, *14*(1), 8353. https://doi.org/10.1038/s41467-023-43458-x

Jones, D. C., Elz, A. E., Hadadianpour, A., Ryu, H., Glass, D. R., & Newell, E. W. (2025). Cell simulation as cell segmentation. *Nature Methods*, *22*(6), 1331–1342. https://doi.org/10.1038/s41592-025-02697-0

Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M., Toth, M., Stockhaus, S., Shrestha, R. B., Rombaut, B., Pollaris, L., Lehner, L., Vöhringer, H., Kats, I., Saeys, Y., Saka, S. K., Huber, W., Gerstung, M., … Stegle, O. (2025). SpatialData: An open and universal data framework for spatial omics. *Nature Methods*, *22*(1), 58–62. https://doi.org/10.1038/s41592-024-02212-x

Marx, V. (2021). Method of the Year: Spatially resolved transcriptomics. *Nature Methods*, *18*(1), 9–14. https://doi.org/10.1038/s41592-020-01033-y

Mehraj, U., Dar, A. H., Wani, N. A., & Mir, M. A. (2021). Tumor microenvironment promotes breast cancer chemoresistance. *Cancer Chemotherapy and Pharmacology*, *87*(2), 147–158. https://doi.org/10.1007/s00280-020-04222-w

Pachitariu, M., Rariden, M., & Stringer, C. (2025). Cellpose-SAM: Superhuman generalization for cellular segmentation. *bioRxiv*. https://doi.org/10.1101/2025.04.28.651001

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., Lotfollahi, M., Richter, S., & Theis, F. J. (2022). Squidpy: A scalable framework for spatial omics analysis. *Nature Methods*, *19*(2), 171–178. https://doi.org/10.1038/s41592-021-01358-2

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100–106. https://doi.org/10.1038/s41592-020-01018-x

Vandereyken, K., Sifrim, A., Thienpont, B., & Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, *24*(8), 494–515. https://doi.org/10.1038/s41576-023-00580-2

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Angerer, P., Bergen, V., Boyeau, P., Büttner, M., Eraslan, G., Fischer, D., Frank, M., Hong, J., Klein, M., Lange, M., Lopez, R., … Community, S. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, *41*(5), 604–606. https://doi.org/10.1038/s41587-023-01733-8

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0