A practical guideline for benchmarking unimodal and multimodal foundation models on spatial transcriptomics data in oncology

Anna C. Wahl-Schaar*, Dasha Valter, Lisa Sikkema, Stefan G. Stark, Ediem Al-Jibury, Tobias Heinen, Liam Gonzalez, Marin Scalbert, Tokuwa Kanno, Zelda Mariet⁺

*Presenting author. ⁺Corresponding author. All authors contributed equally. Correspondence to contact@bioptimus.com.

AACR

LB159

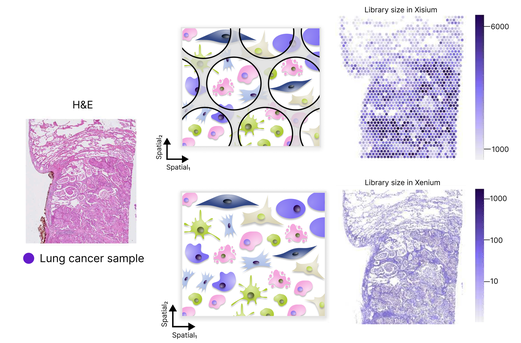
BIOPTIMUS

ABSTRACT

Understanding the molecular and morphological structure of the tumor microenvironment is increasingly approached by the use of foundation models (FMs): large-scale, complex machine learning architectures trained on H&E, spatial omics, and/or molecular data. In parallel, data generation and experimental assays measuring molecular features in their spatial context are growing in size, resolution, and scale, often no longer limited to only a few hundred genes, but multiple thousand features measured at subcellular resolution. This combination of scaled data generation and complex, often black-box, models requires well-defined benchmarks and transparent baselines to guide practitioners in deciding which approach(es) will best suit a specific goal with specific data. However, existing benchmarks are designed in a computationally driven manner and focus only on a coarse-grained set of labels and genes. Although such benchmarks can give a good indication of overall performance, they often fail to capture different disease types and downstream use-cases which in practice will require distinct data types, outputs, and evaluation metrics. They also fail to capture the higher-level aim of foundation models: enabling novel biological findings. Finally, reproducing benchmark results on internal or novel datasets is often limited by complex computational frameworks.

This work provides recommendations for building meaningful benchmarks for foundation models on spatial transcriptomics data, as defined by the combination of (a) simple, transparent baselines, (b) diverse datasets spanning various indications and data types, and (c) biologically relevant downstream tasks with interpretable metrics. We begin by providing a complete overview of current benchmarking efforts leveraging spatial transcriptomics data and their design choices as well as limitations. To enhance existing benchmarks, we introduce a new set of tasks and metrics designed on common cell and gene signatures present in oncology, and allow for different processing choices designed for subcellular spatial transcriptomics data. For each task, we also define a simple, memory-efficient baseline in order to put the performance of current and upcoming foundation models into perspective. We conclude by discussing useful guidelines for defining future benchmarking datasets for foundation models. The novel tasks and benchmarking frameworks introduced in this work provide a crucial toolset to researchers seeking to accurately assess the evolving landscape of uni- and multimodal foundation models for spatial omics.

The trade-off between spatial resolution and large plex measurements

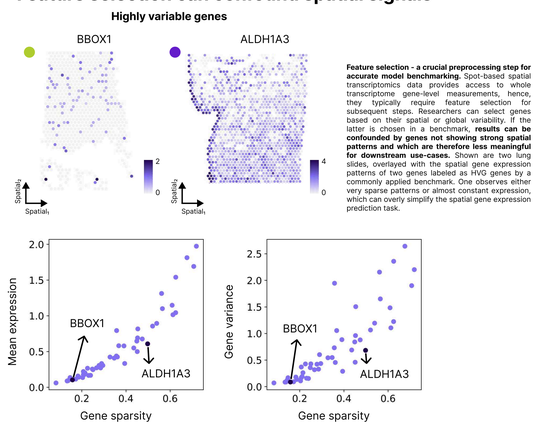


The resolution trade-off in spatial transcriptomics data. Researchers face an important choice when designing spatial transcriptomics experiments and are typically confronted with a trade-off between resolution with respect to capture size, spatial resolution and gene coverage and throughput. Depending on the goal of the analysis, each assay comes with its own pros and cons.

The resolution trade-off is often ignored in benchmarking frameworks and datasets are harmonized into a common data structure ignoring the technological differences present in both. Aggregation strategies vary between frameworks, e.g. Xenium sometimes is binned to hexagonal or regular grids of sizes chosen by the original method.

Shown is a Post-Xenium H&E measuring a patient diagnosed with lung adenocarcinoma and the associated spatial plot of counts binned into a Visium-like hexagonal structure at spot diameter of 55 micrometers, called Xisium, (top panel) versus the cellular resolution (bottom panel). Both panels are independently colored by the library size of the spot or the cell. **Depending on the preprocessing choice of SpT technologies, spot profiles are much stronger confounded by library size effects.** This technical artefact must be considered during downstream analysis.

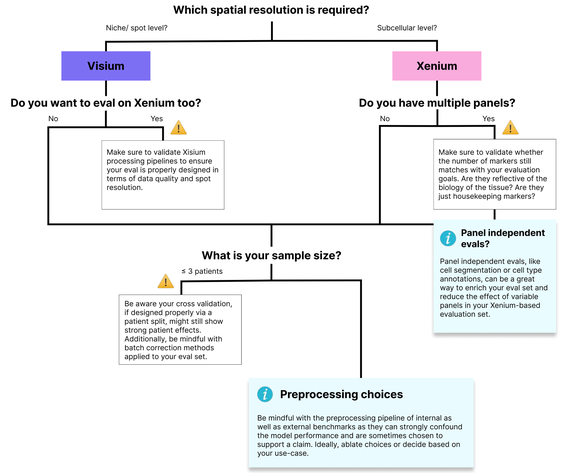
Feature selection can confound spatial signals



Feature selection - a crucial preprocessing step for accurate model benchmarking. Spot-based spatial transcriptomics data provides access to whole transcriptome gene-level measurements, hence, they typically require feature selection for subsequent steps. Researchers can select genes based on their spatial or global variability. If the latter is chosen in a benchmark, **results can be confounded by genes not showing strong spatial patterns and which are therefore less meaningful for downstream use-cases.** Shown are two lung slides, overlayed with the spatial gene expression patterns of two genes labeled as HVG genes by a commonly applied benchmark. One observes either very sparse patterns or almost constant expression, which can overly simplify the spatial gene expression prediction task.

Spatial gene sparsity as an additional source of feature selection. We propose to additionally consider spatial gene expression sparsity, as some genes might only be present in very few spots and could be outliers or artefacts captured by the assay. Conversely, when assessing genes which are expressed in every spot, one should closely inspect their spatial variability to ensure meaningful signals during model evaluations.

A checklist for evaluation and benchmark design



The benchmarking landscape for predicting spatial transcriptomics

Benchmark	Benchmark name	Input modality	Primary metric
Jaume, G. et al. 2024	HEST1K	HnE	PCC (50 HVG)
Wang, C. et al. 2025		HnE	PCC (HVG, 20 SVG), MI, SSIM, AUC
Gindra, R.H. et al. 2025	HESCAPE	HnE, scRNAseq	PCC (50 HVG as of HEST1K)

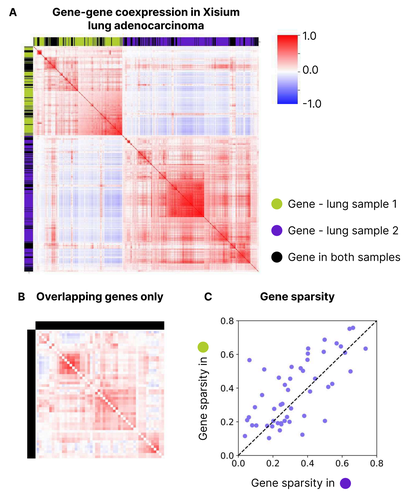
Method paper with benchmarking aspect	Method name	Input modality at inference time	
Comiter, C. et al. 2025	SCHAF	HnE, (scRNAseq)	PCC (all genes), Accuracy (cell type annotations)
Konchev, K. et al. 2025	DeepSpot	HnE	PCC (top N 50-5000)
Zhang, N. et al. 2025	SpaFoundation	HnE	PCC, SCC, MSE, RMSE
Wang, Y. et al. 2025	FmH2ST	HnE	PCC, Spearman's rank correlation coefficient, MSE, MAE, ARI

The benchmarking landscape for predicting spatial transcriptomics. A selection of currently widely used benchmarking and method papers evaluating the performance of FMs for the task of predicting spatial gene expression profiles. Due to limitations in data availability, **presumably independent benchmarks often depend on the same data corpus.** Compared to other fields, evidence of superior FM performance remains relatively sparse. Additionally, the lack of diversity in reported metrics currently hinders an objective assessment of the true predictive value and practical applicability of the competing methods.

Different targeted SpT panels provide varying insights

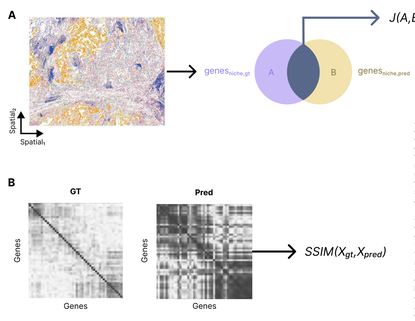
Task name	Panel name	Markers in panel	Markers retained after join	Shared markers across all
Pancreatic adenocarcinoma	Xenium V1 hPancreas Cancer with Add on	474	98	42
	Xenium V1 Human Pancreas	377		
	Xenium V1 Human Ductal Adenocarcinoma	380		
Lung adenocarcinoma	Xenium V1 Human Lung Cancer	377	198	
	Xenium V1 Human Lung Cancer Addon	480		
Colon adenocarcinoma	Xenium V1 hColon Cancer Add on FFPE	425	115	
	Xenium V1 Human Colorectal Cancer Add on FFPE	480		

Targeted panels provide different insights per slide. Targeted spatial transcriptomics strongly depends on panel design and different commercially available panels do not necessarily capture the same biology on the slide. When merging samples measured with different panels, one significantly reduces the gene set available for evaluations and benchmarking. Additionally, even if a gene is present in multiple panels, the probe coordinates and probe sequences can be different. **The limitations should be carefully considered when formulating internal evaluations and assessing external independent benchmarks.** The colored dots match the co-expression heatmap computed for this task.



Limited data availability can confound performance when applying ill-posed preprocessing pipelines. Some widely applied benchmarking frameworks propose to curate tasks independent of panel, but only with a focus on e.g. cancer indication. However, panels might introduce strong batch effects and genes measured in one versus the other can show very diverging patterns. When testing models on these tasks, **researchers should carefully consider data preprocessing choices without blindly trusting processed data sources.**

Additional biologically informed metric beyond correlation



A metric to understand the biological signals across niches in SpT data. Niches and regions are a common way of annotating spatially resolved omics data. Current evaluations typically do not assess whether niche signatures are recovered by a model's predictions. **We propose to calculate niche signatures, both in ground truth (gt) SpT data as well as predicted SpT profiles**, and compare their signatures using set-based metrics similar to *Jaccard index*.

Structural similarity as a metric to assess gene-gene co-expression stability. Global cell-, spot- or gene-wise metrics typically do not account for biological signals. Genes might exhibit relatively low variance both globally and in a spatially-aware manner, but nevertheless they can show very strong co-expression patterns between groups of genes. **We propose to additionally calculate image-based metrics on the obtained gene-gene co-expression profiles**, e.g. *structural similarity*. Additionally, one can pair these metrics with known communication pathways.

CONCLUSION

The growing abundance of spatial transcriptomics data provides a meaningful playground for evaluation of uni- and multimodal foundation models. In principle, if designed well, foundation models can significantly reduce cost and time required to understand the tumor microenvironment. However, in order to fully leverage the predictive power of these models, and put them into perspective, one needs well designed and transparent benchmarks.

This work reveals that current benchmarking efforts are often constrained by limited data diversity and narrow evaluation metrics, and frequently overlook technology-specific biases. Consequently, these frameworks likely fail to provide a robust resolution of model performance across the diverse landscape of clinical indications. To address this, **we have proposed a guide for designing well-phrased benchmarking frameworks which will help researchers to formulate their own internal benchmarks**, as well as better assess the limitations of prior work. We showed the limitations of using targeted spatial transcriptomics data with different panels and argued for stringent data preprocessing, careful consideration for feature selection, and cross validation designs. We proposed additional metrics beyond Pearson and Spearman correlation, which can provide more fine-grained insights.

We believe that our work will help researchers and scientists define stronger internal and external benchmarks and gain the right toolset to re-assess the usability of existing work.

Selected reference

- Jaume, G., et al. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis. Advances in Neural Information Processing Systems, December 2024.
- Wang, C., Chan, A.S., Fu, X. et al. Benchmarking the translational potential of spatial gene expression prediction from histology. Nat Commun 16, 1544 (2025).
- Gindra, R. H., et al. A Large-Scale Benchmark of Cross-Modal Learning for Histology and Gene Expression in Spatial Transcriptomics. arXiv (2025).
- Comiter, C., et al. Inference of single cell profiles from histology stains with the Single Cell omics from Histology Analysis Framework (SCHAF). bioRxiv (2025).
- Nonchev, K., et al. DeepSpot: Leveraging Spatial Context for Enhanced Spatial Transcriptomics Prediction from H&E Images. medRxiv (2025).
- Zhang, N., et al. Inferring spatial gene expression from tissue images using large-scale histology foundation model with SpaFoundation. bioRxiv (2025).
- Wang, Y., et al. FmH2ST: foundation model-based spatial transcriptomics generation from histological images. Nucleic Acids Research, 53(17) (2025).
- Marco Salas, S. et al. Optimizing Xenium In Situ data utility by quality assessment and best-practice analysis workflows. Nat Methods 22, 813–823 (2025).
- Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023).
- Tejada-Lapuerta, A., Schaar, A.C., Gutgesell, R. et al. Nicheformer: a foundation model for single-cell and spatial omics. Nat Methods 22, 2525–2538 (2025).

Stay tuned for our most recent releases & initiatives!

M-Optimus

STELA

Bioptimus booth number: 3654

Contact details:
contact@bioptimus.com

AACR Annual Meeting 2026 - San Diego