nature portfolio

Corresponding author(s): Jin Liu

Last updated by author(s): Mar 11, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All datasets utilized in this study are publicly accessible.
Data analysis	All analyses were performed using Python (v3.10). The CoxFormer methods were implemented in an open-source, publicly available Python package that is available at https://github.com/yyyancy/CoxFormer. Code for reproducing the analysis can be found at https://github.com/yyyancy/CoxFormer. General data processing and statistical analyses were conducted using NumPy, SciPy, scikit-learn and pandas. Figures were generated using Matplotlib and Seaborn. Additional Python packages were used for running numerical studies: Highly variable gene (HVG) selection and differential expression analysis were performed using Scanpy (v1.9.8); Gene set enrichment analyses were performed using GSEApy (v1.1.8); Cell–cell communication analyses were performed using LIANA (v1.6.1). For benchmarking transcriptomics prediction, we compared CoxFormer with eight existing spatial transcriptomics integration or imputation methods: gimVI (version 0.8.0), SpaGE (no version), Tangram (version 1.0.0), Seurat (version 5.3.0), SpaOTsc (version 0.2), LIGER (version 0.5.0), novoSpaRc (version 0.4.3), and stPlus (version 0.0.6). Each baseline method was implemented following the official documentation or publicly available repositories with recommended parameter settings to ensure fair comparison.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets utilized in this study are publicly accessible:

COXPRESdb dataset
The COXPRESdb microarray co-expression dataset (Microarray_m.c7 version) can be downloaded from https://coxpresdb.jp/download/.

Human Cell Atlas scRNA-seq dataset
The Human Cell Atlas single-cell RNA sequencing dataset is available at https://explore.data.humancellatlas.org/projects.

Gene classification dataset
The gene classification benchmark datasets are available at https://huggingface.co/datasets/ctheodoris/Genecorpus-30M/tree/main/example_input_files/gene_classification.

Cell-level scRNA-seq datasets
The scRNA-seq datasets used for the cell-level tasks include Diffuse Large B-cell Lymphoma FFPE (DLBL), Breast Cancer FFPE (BC), and Lung Cancer FFPE (LC), which are accessible from the CZ CELLxGENE collection at https://cellxgene.cziscience.com/collections/bd552f76-1f1b-43a3-b9ee-0aace57e90d6. The Bone Marrow (BM) dataset can be downloaded using scvelo.datasets.bonemarrow() in the Python package scvelo.

10x Visium human breast cancer spatial transcriptomics dataset
The 10x Visium human breast cancer spatial transcriptomics dataset (HBC1–HBC6) is available from Zenodo at https://zenodo.org/record/4739739, corresponding to samples CID4535 (HBC1), CID44971 (HBC2), CID4465 (HBC3), CID4290 (HBC4), 1160920F (HBC5), and CID3586 (HBC6).

10x Chromium human breast cancer scRNA-seq dataset
The 10x Chromium human breast cancer scRNA-seq dataset is available from GEO under accession GSE176078 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078.

Human brain spatial ATAC–RNA-seq dataset
The spatial ATAC–RNA-seq human brain hippocampal spatial transcriptomics and epigenomics dataset is available at https://cells.ucsc.edu/?ds=brain-spatial-omics.

10x Xenium human skin melanoma dataset (Section A)
The Xenium human skin melanoma dataset (Section A) can be accessed from 10x Genomics at https://www.10xgenomics.com/datasets/human-skin-preview-data-xenium-human-skin-gene-expression-panel-add-on-1-standard.

10x Xenium human skin melanoma dataset (Section B)
The Xenium human skin melanoma dataset (Section B) can be accessed from 10x Genomics at https://www.10xgenomics.com/datasets/human-skin-preview-data-xenium-human-skin-gene-expression-panel-1-standard.

10x Visium colorectal cancer liver metastasis spatial transcriptomics dataset
The colorectal cancer liver metastasis spatial transcriptomics dataset (CRC1–CRC2 and LM1–LM2) is available from scCRLM at http://www.cancerdiversity.asia/scCRLM, corresponding to samples ST-P1, colon (CRC1), ST-P4, colon (CRC2), ST-P2, liver (LM1), and ST-P4, liver (LM2).

Normal human colon scRNA-seq dataset
The Illumina NextSeq 500 scRNA-seq dataset from normal human colon is available from GEO under accession GSM7290762 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7290762.

Normal human liver scRNA-seq dataset
The Illumina NextSeq 500 scRNA-seq dataset from normal human liver is available from GEO under accession GSM7290760 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7290760.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA

Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by the sizes of the publicly available datasets used in this study. No statistical methods were used to predetermine sample size.
Data exclusions	For the human breast cancer Visium datasets (HBC1–HBC6), spatial transcriptomics profiles were filtered through a gene quality control procedure before downstream analysis. Specifically, HBC1 contained 1,125 spatial spots and 16,928 genes before filtering and 10,432 genes after filtering. HBC2 contained 1,160 spatial spots and 17,479 genes before filtering and 10,368 genes after filtering. HBC3 contained 1,211 spatial spots and 17,365 genes before filtering and 10,678 genes after filtering. HBC4 contained 2,426 spatial spots and 17,508 genes before filtering and 8,948 genes after filtering. HBC5 contained 4,895 spatial spots and 20,002 genes before filtering and 9,326 genes after filtering. HBC6 contained 4,784 spatial spots and 19,451 genes before filtering and 7,896 genes after filtering. From the filtered genes of each dataset, 1,000 highly variable genes were selected for downstream analysis. For the human brain hippocampal spatial transcriptomics dataset, the data contained 2,500 spatial spots and 21,179 genes before quality control and 14,463 genes after filtering. From these genes, 1,000 highly variable genes were selected for downstream analysis. For the colorectal cancer liver metastasis spatial transcriptomics datasets (CRC1–CRC2 and LM1–LM2), CRC1 contained 3,313 spatial spots and 36,601 genes before filtering and 11,000 genes after filtering (including 2,018 housekeeping genes). CRC2 contained 3,902 spatial spots and 36,601 genes before filtering and 8,782 genes after filtering (including 1,959 housekeeping genes). LM1 contained 4,658 spatial spots and 36,601 genes before filtering and 10,569 genes after filtering (including 2,010 housekeeping genes). LM2 contained 3,721 spatial spots and 36,601 genes before filtering and 6,063 genes after filtering (including 1,706 housekeeping genes). All filtering procedures removed low-quality or lowly expressed genes prior to model training.
Replication	We applied CoxFormer to each dataset. The results presented in the paper can be replicated with code available at our website https://github.com/yyyancy/CoxFormer.
Randomization	We did not perform any randomized experiments that involves assigning individuals to groups.
Blinding	We did not perform any experiments that involves assigning individuals to groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA