nature methods

Analysis

https://doi.org/10.1038/s41592-024-02215-8

Benchmarking spatial clustering methods with spatially resolved transcriptomics data

Received: 13 March 2023

Accepted: 16 February 2024

Published online: 15 March 2024

Check for updates

Zhiyuan Yuan[1,2,7] ✉, Fangyuan Zhao[3,4,7], Senlin Lin[3,4], Yu Zhao[5], Jianhua Yao[5], Yan Cui[1,2,6], Xiao-Yong Zhang[1] & Yi Zhao[3,4] ✉

Spatial clustering, which shares an analogy with single-cell clustering, has expanded the scope of tissue physiology studies from cell-centroid to structure-centroid with spatially resolved transcriptomics (SRT) data. Computational methods have undergone remarkable development in recent years, but a comprehensive benchmark study is still lacking. Here we present a benchmark study of 13 computational methods on 34 SRT data (7 datasets). The performance was evaluated on the basis of accuracy, spatial continuity, marker genes detection, scalability, and robustness. We found existing methods were complementary in terms of their performance and functionality, and we provide guidance for selecting appropriate methods for given scenarios. On testing additional 22 challenging datasets, we identified challenges in identifying noncontinuous spatial domains and limitations of existing methods, highlighting their inadequacies in handling recent large-scale tasks. Furthermore, with 145 simulated data, we examined the robustness of these methods against four different factors, and assessed the impact of pre- and postprocessing approaches. Our study offers a comprehensive evaluation of existing spatial clustering methods with SRT data, paving the way for future advancements in this rapidly evolving field.

Advancements in spatially resolved transcriptomics (SRT) enable the multiplexed spatial mapping of gene expression, allowing researchers to move beyond cell clustering to identify higher-order tissue structures, or spatial domains, through the provision of additional spatial information[1–3]. Identifying spatial domains by spatial clustering has become a standard initial step in constructing spatial atlas[4–8] and has proven to be crucial in visualizing tissue anatomy[9], inferring tissue spatial continuity[10,11], detecting domain-specific marker genes[12,13], mining spatial signatures of development and disease[14,15], and identifying domain-dependent molecular regulatory networks[16,17] (Supplementary Note 1).

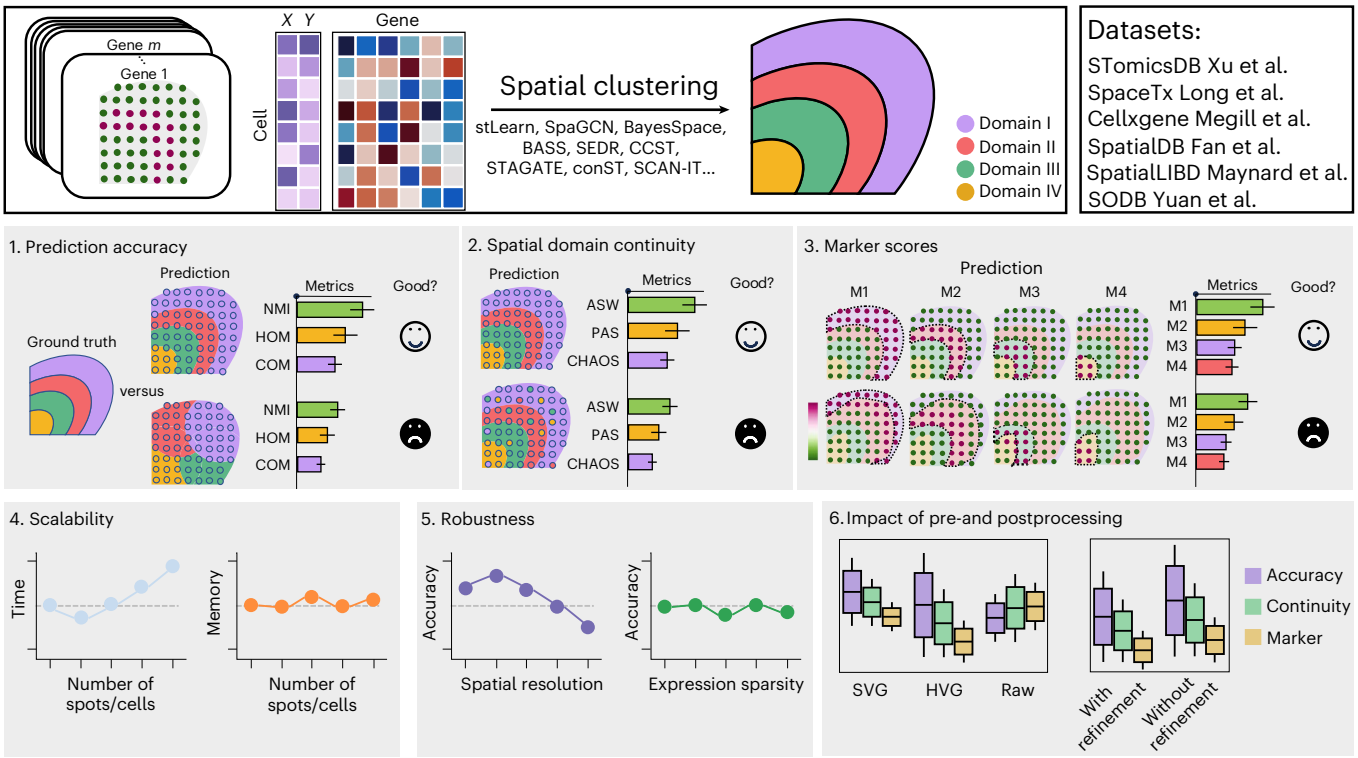
Despite the availability of computational methods based on probabilistic graphical models and graph neural networks (GNNs) for identifying spatial domains recent years[18,19], the lack of consistency and comprehensiveness in the datasets and metrics used poses substantial challenges. These difficulties arise from the rapid advancements in spatial technologies, the limited evaluation metrics used in certain applications, and the reliance on datasets obtained from specific technologies and tissues that have been generated by certain laboratories (Supplementary Note 2).

Although benchmarking efforts have been made for spatially resolved transcriptomics data, particularly in relation to cell type[20], a comprehensive benchmark study specifically targeting spatial clustering methods designed for identifying spatial domains is still needed (Extended Data Fig. 1). In this Analysis, we considered ten metrics of

[1]Center for Medical Research and Innovation, Shanghai Pudong Hospital, Fudan University Pudong Medical Center, Fudan University, Shanghai, China. [2]Institute of Science and Technology for Brain-Inspired Intelligence; MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence; MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China. [3]Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [4]University of Chinese Academy of Sciences, Beijing, China. [5]Tencent AI Lab, Shenzhen, China. [6]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan. [7]These authors contributed equally: Zhiyuan Yuan, Fangyuan Zhao. ✉e-mail: zhiyuan@fudan.edu.cn; biozy@ict.ac.cn

https://doi.org/10.1038/s41592-024-02215-8

a



b

Spatial technology	Data number	Size/ radius	Spots	Genes	Sparsity of examined data
10× Visium	12	55 μm	3,844	33,538	0.94
Stereo-Seq	9	25 μm	5,913	24,238	0.87
BaristaSeq	3	≤1 cell	1,690	79	0.81
MERFISH	5	≤1 cell	5,557	155	0.58
osmFISH	1	≤1 cell	4,839	33	0.31
STARmap	3	≤1 cell	1,053	166	0.70
STARmap*	1	≤1 cell	1,207	1,020	0.76
Simulated_1	10	≤1 cell	3,639	33,538	0.98
Simulated_2	10	≤1 cell	1,136	33	0.10
Simulated_3	5	≤1 cell	4,839	33	0.31
Simulated_4	120	≤1 cell	3,667	2,136	0.92

Fig. 1 | Pipeline and data. a, The pipeline of the benchmark study. **b**, The datasets used in this study. Information such as number of subdatasets, spatial resolution, number of spots/cells, number of genes, and gene expression matrix sparsity is listed. Note that the lengths of the bars are proportional to mean values, and the number beside the bars are median values. Detailed information for the data source is presented in Methods. Data are presented on the basis of mean and 95% confidence intervals. Number of data are shown in 'Data number' column.

four categories (prediction accuracy, spatial domain continuity, the ability to detect domain marker genes, and scalability; Methods). For the purpose of clarity in this paper, we used 'accuracy' in a broader sense than its traditional statistical usage, and 'data' to refer to spatial transcriptomics data from a single slice, while a 'dataset' refers to a collection of data from the same publication, sharing the same technology.

Our first step involved a comprehensive analysis of spatial clustering methods within the context of spatially resolved transcriptomics. We benchmarked 13 computational methods across 34 data (7 datasets) from various spatial technologies (Fig. 1, Methods and Supplementary Table 1). Data were selected on the basis of the variety of data types and the availability of annotations. Our study found no 'one-size-fits-all' method worked well on every data. We provided user guidance for selecting the optimal spatial clustering methods based on data characteristics and introduced a website interface (http://sdmbench.drai.cn/) for benchmarking new methods against existing ones.

Second, we identified limitations within current methods. These limitations became apparent during the testing of various spatial clustering methods on additional 22 data (Methods and Supplementary Table 1) containing small and noncontinuous tissue domains, and during multislice analysis on another large-scale dataset containing 31 tissue slices (Methods and Supplementary Table 1). We proposed a 'divide and conquer' strategy to make the latter challenging task effectively solvable.

Lastly, we tested the impact of robustness and other factors. With 145 simulated data in total, we evaluated the impact of several factors on the performance of these methods, including gene expression matrix sparsity, spatial resolution, the number of genes and the level of noise (Methods). We also tested the impact of pre- and postprocessing steps on different methods' performance.

Overall, this work represents a profound contribution to the field of spatially resolved transcriptomics, providing a comprehensive

evaluation framework for spatial clustering methods and facilitating their application to a wide range of datasets. We anticipate that our findings will be of interest to researchers in the field and have significant implications for advancing the development of more effective and efficient spatial clustering methods in increasingly complex use cases.

Results

Methods and datasets

To evaluate the performance of various spatial clustering methods (Fig. 1a), the primary criterion for dataset selection was the availability of spatial domain annotations as ground truth. We also evaluated the quality of those annotations (Supplementary Note 3 and Supplementary Figs. 1–3). To this end, various extant spatial transcriptomics databases, including STomicsDB[21], SpaceTx[22], Cellxgene[23], SpatialDB[24], SpatialLIBD[25] and SODB[26], were extensively examined. This led to the acquisition of 34 real data from a variety of spatial technologies, including 10x Visium[25], Stereo-Seq[27], BaristaSeq[28], MERFISH[29], osmFISH[30], STARmap[31] and STARmap*[31] (1k gene version of STARmap) (Methods, Fig. 1b and Supplementary Table 1). As these data were produced using distinct spatial technologies, they exhibit different data characteristics and collectively span a wide range of potential spatial transcriptomics data types, which allows users of different spatial technologies to benefit from the study. The reliability of these datasets and their annotations has made them a frequent choice for methods development[12,32].

A total of 13 computational methods (Methods), including 11 spatial clustering methods and 2 nonspatial clustering methods, were considered in this study. The benchmarked spatial methods were mostly developed after 2021, representing the most recent advancements. The nonspatial methods employed were Louvain[33] and Leiden[34]. The spatial methods include SpaGCN[12], BayesSpace[35], stLearn[36], SEDR[11], CCST[13], SCAN-IT[37], STAGATE[38], SpaceFlow[32], conST[39], BASS[40] and GraphST[41]. Although Louvain and Leiden are both designed for nonspatial single-cell data, following previous studies[12,35,36], we included them as the baseline to illustrate the effect of considering spatial information.

As suggested by previous benchmark studies[42], the diversities of spatial technologies and computational methods are both necessary to achieve the primary goal of this work, which is to guide biologists in choosing optimal methods and developers in improving current state-of-the-art. Biologists need to know the optimal method for their datasets, whether they were generated by existing or new spatial technologies. Of the existing spatial technologies, 10x Visium and MERFISH, which were both involved in this study, occupy a substantial portions of published datasets and will probably be more popular due to their commercialization[43]. For datasets generated by new spatial technologies, this benchmark study can also provide useful information by inferring the optimal methods on the basis of similarities in data characteristics.

The diverse range of computational methods is also necessary, even if some methods are specifically designed for particular spatial technologies. Many potential users without programming skills may not know or care about the underlying principles and designs of each method or the statistical nature of data generated from different spatial technologies. Therefore, we provide unadorned results of different methods on various data types as the most straightforward way to help a wide range of potential users and readers. Reporting the multi-aspect performance of diverse methods on diverse datasets would also motivate method developers to identify important limitations that exist in current methods on specific datasets. In this study, we identified some limitations in current methods and demonstrated how these could be tentatively resolved by the combination of our recommended method and existing tools. To better contribute to the field, we provided both Python package and website for new methods benchmarking (Supplementary Note 4).

Benchmarking analysis on 10x Visium dataset

The benchmarking pipeline (Fig. 1a) is illustrated using the DLPFC 10x Visium dataset[25] as an example (Fig. 2a,b). This dataset contains 12 tissue slices from the human dorsolateral prefrontal cortex (DLPFC), with each slice containing more than 20,000 genes at a transcriptome scale, with a median of 3,844 spots (Methods). Additionally, this dataset is frequently employed to evaluate almost every spatial clustering method. Initially, the performance of spatial clustering on slice#151673 (Data9) was analyzed by plotting the predicted labels in tissue space (Fig. 2a). Most methods exhibited laminar patterns as expected when compared to the ground truth (Fig. 2a), particularly BayesSpace, and two recent methods, BASS and GraphST.

Clustering accuracy was quantified using normalized mutual information (NMI), the most widely used accuracy metric in the field. The top-performing methods for slice#151673 (Data9) (Fig. 2a) were BASS (NMI 0.71), GraphST (NMI 0.697), and BayesSpace (NMI 0.688). When the performance was aggregated across all 12 slices, a highly consistent top method was observed (Fig. 2b), with GraphST (median NMI 0.621) and BASS (median NMI 0.612) being the most accurate, followed by SCAN-IT (median NMI 0.599) and BayesSpace (median NMI 0.598) with similar accuracies. As expected, Leiden (median NMI 0.246) and Louvain (median NMI 0.240) exhibited the least performance since they did not incorporate spatial information and were often used as baseline methods by spatial clustering papers to demonstrate the benefits of using spatial information.

To provide more comprehensive evaluations, other metrics were utilized (Methods). For accuracy, the homogeneity score (HOM)[44] was incorporated, which incentivizes methods that generate subdomains within the ground truth domains. The completeness score (COM)[44] was also used, which encourages methods that produce larger domains that encompass several true domains. Continuity is another important measure for spatial clustering as it encourages predicted labels to exhibit spatial coherence and clear domain-domain interfaces. To assess continuity of predicted spatial domains, three widely used metrics were employed, namely CHAOS[45], percentage of abnormal spots (PAS)[45] and average silhouette width (ASW)[46]. Identification of domain-specific marker genes is a crucial biological application, and the ability of each method to identify high-quality domain-specific marker genes was further evaluated. Spatial autocorrelation is the most widely accepted metric for evaluating the quality of domain-specific marker genes[12], and Moran's *I* (ref. 47) and Geary's *C* (ref. 48) (Methods) were computed and averaged across top marker genes of each method to report the performance.

Although NMI is the most important metric as demonstrated in previous studies[12,32,35], other metrics can reflect complementary aspects of method performance and are valuable for users' reference. When selecting a method, users can jointly consider multiple metrics based on their specific needs. For example, if finer tissue structures are required, methods with higher HOM, such as SpaceFlow, might be preferred (Fig. 2b). However, SpaceFlow's low NMI and tail-ranked marker score may make it a suboptimal choice for 10x Visium datasets (SpaceFlow's strength is mainly in imaging-based spatial datasets, see next sections). For another example, if smoother boundaries are needed, BASS, with a better continuity score, may be the better choice, even though BASS and GraphST exhibit similarly good NMI scores (Fig. 2b).

Benchmarking analysis on MERFISH dataset

MERFISH is a highly utilized imaging-based spatially resolved transcriptomics (SRT)[7,29,49] technology, and its potential future applications may become more widespread due to commercialization. We analyzed a MERFISH dataset of the mouse hypothalamic preoptic region[50], which contains five annotated slices with a median of 5,557 cells (Fig. 1b and Methods). This prediction is particularly challenging given the complex tissue structure and domains with heterogeneous shapes and adjacency (Fig. 2c). Three available methods in the 10x Visium dataset, that is, SpaGCN (HE), BayesSpace and stLearn, were not applicable to this dataset (Methods).

https://doi.org/10.1038/s41592-024-02215-8

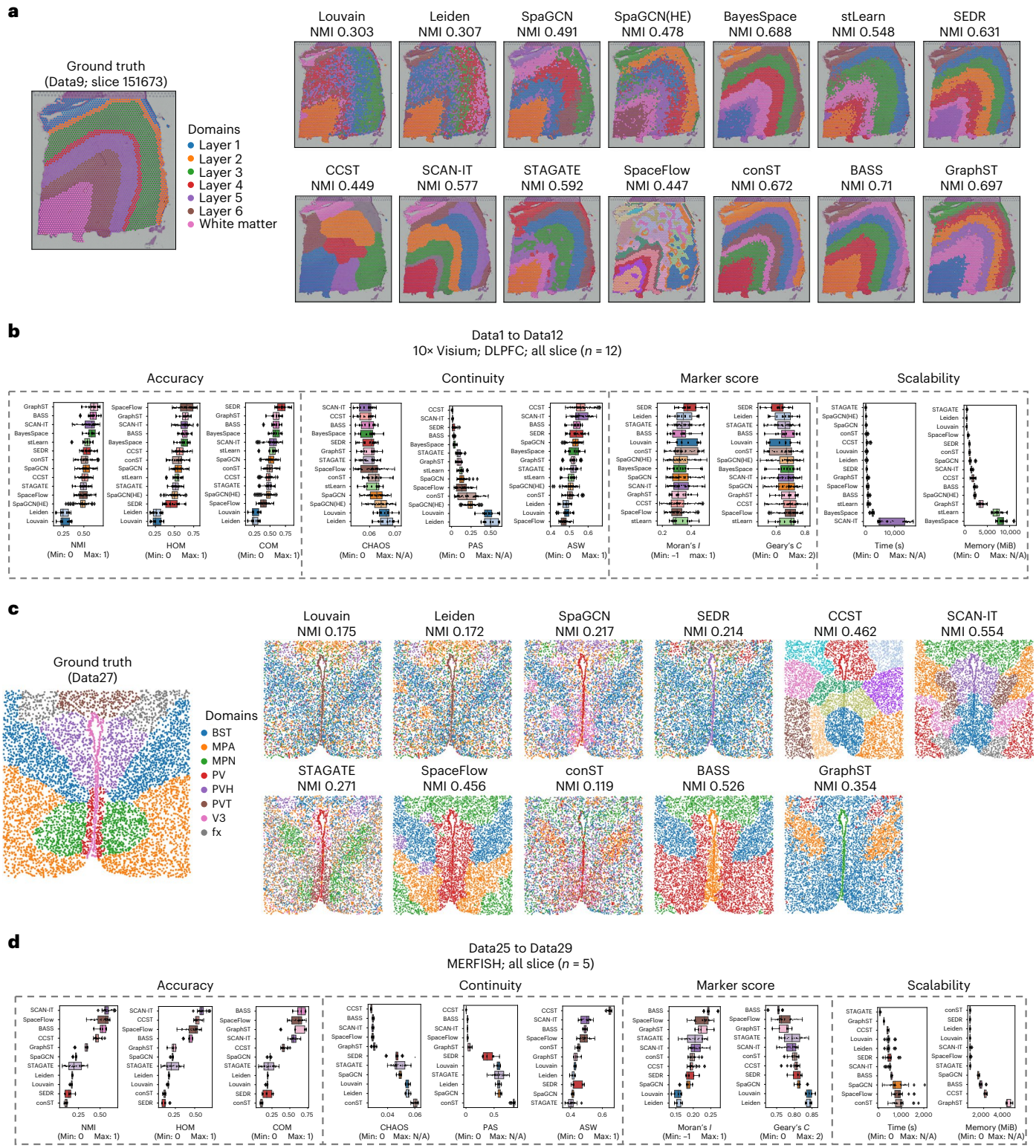


Fig. 2 | Evaluation on 10x Visium and MERFISH. a, The ground truth and methods' output on a representative slice of 10x Visium dataset. **b**, Quantitative evaluations were recorded on all 12 data of 10x Visium. On each data, each method was repeated for ten times. For each metric, the compared methods are arranged in order from the best performance to worst. Center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range. $N = 10$ independent runs for 12 data. **c**, The ground truth and methods' output on a representative slice of MERFISH dataset. BST, bed nuclei of the strata terminalis; V3, third ventricle; PV, periventricular hypothalamic nucleus; PVT, paraventricular nucleus of the thalamus; MPN, medial preoptic nucleus; fx, columns of the fornix; PVH, paraventricular hypothalamic nucleus; MPA, medial preoptic area. **d**, Quantitative evaluation was recorded on all five data of MERFISH. On each data, each method was repeated for ten times. For each metric, the compared methods are arranged in order from the best performance to worst. Center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range. $N = 10$ independent runs for five data.

Quantitative metrics of all available methods on the MERFISH dataset (Fig. 2d) revealed significant differences compared to their performance on the 10x Visium dataset (Fig. 2b). The best-performing method, GraphST, in the 10x Visium dataset (ranked 1 in NMI) did not rank as the top method in the MERFISH dataset (ranked 5 in NMI). The continuity of CCST was the top among the 11 methods for MERFISH, and the accuracy of CCST for the MERFISH dataset improved, particularly in NMI, compared to that for the 10x Visium dataset, indicating that CCST was better suited for imaging-based datasets (Fig. 2d). BASS, SpaceFlow and SCAN-IT were the methods that could best capture domain-specific marker genes and had both top accuracy and continuity, making them the recommended methods to choose for MERFISH data (Fig. 2d). We also analyzed the performance variance in the 10x Visium and MERFISH dataset (Supplementary Note 5). Benchmarking analyses on other spatial data types are also available (Supplementary Figs. 4–8 and Supplementary Note 6).

Overall performance

Through the analysis of methods performance across different data, we have observed interesting correlation patterns as determined by the Spearman correlation coefficient, underlining the relative ordering of their performance (Fig. 3a, the underlying metrics values can be found in Extended Data Fig. 2). First, we found that, within the same spatial technology, a strong Spearman correlation emerged when multiple data samples originated from the same individual and shared identical tissue structure. A notable example of this is the high consistency in the relative performance of all methods applied on three BaristaSeq datasets (Fig. 3a, blue box). Second, when tissue structure is controlled, methods performance may be influenced by the use of different technologies. Evidence for this is seen when comparing the performance of methods on the 10x Visium dataset with that of the BaristaSeq dataset. Sharing similar tissue structures (brain cortex region, Supplementary Table 1), the two datasets, generated by different spatial technologies, yielded different performance outcomes (Spearman's correlation is 0.33, Fig. 3b). Third, when spatial technology is controlled, the origin of the data samples (that is, different donors) can also influence method performance. An example is on the 10x Visium dataset, the significantly higher correlations (Supplementary Fig. 9, $P = 4.65 \times 10^{-4}$) when applying methods on data from the same donor than from the different donors (correlation values from the same donor are highlighted in red box in Fig. 3a).

Based on the correlated performance of methods across spatial technologies, we have summarized the performance of all methods by partitioning the datasets into two groups. The first group comprises datasets generated by 10x Visium, and the second group comprises datasets generated by imaging-based datasets (MERFISH, osmFISH, BaristaSeq, STARmap and STARmap*). The performances displayed (Fig. 4 and Supplementary Fig. 10) were based on a rank-based score (Methods), as done in previous benchmark analyses[42], so higher scores are better. Users can easily derive their recommendations on the basis of their circumstances.

The overall evaluation demonstrated the complementary nature of all methods in terms of their performance and functionality. BASS and SCAN-IT displayed top performance in both groups of datasets (Fig. 4a,b), indicating the best generalizability, although their scalability was relatively lower than that of other methods. Users may want to choose more efficient methods, such as STAGATE, conST or SEDR, if some compromise in prediction accuracy is acceptable. STAGATE achieved the top scalability score in all cases, but could not achieve top accuracy. Some methods exhibited technology-biased performance. For example, BayesSpace performed particularly well on 10x Visium dataset, but cannot be applied to imaging-based datasets and has a longer running time. For imaging-based datasets (Fig. 4b), SpaceFlow and CCST generated both good accuracy and top scalability, while their accuracy ranking was not at the top in the 10x Visium dataset. We have provided user guidance for choosing methods in Extended Data Fig. 3. Note that the criterion is based on accuracy. If other aspects should be considered, users can refer to Fig. 4 for more comprehensive recommendations. All the benchmarking results are available at http://sdmbench.drai.cn/ (Supplementary Figs. 11–13).

One can observe a major proportion of Python methods compared to R (Fig. 4a,b). The Python methods are mostly based on the deep learning approach, which has benefited from the advancement of the deep learning ecosystem, such as PyTorch (https://pytorch.org/) and PyG (https://www.pyg.org/). Most deep learning methods can generate a 'context-aware representation' which is a type of cellular representation that encodes both the gene expression and spatial information, along with the spatial domain labels. This context-aware representation of cells can be freely incorporated with existing single-cell tools for spatially related tasks. For example, the pseudo-spatiotemporal can be obtained by applying trajectory-based single-cell methods on the learned context-aware representation to unravel spatiotemporal patterns of cells[32].

Limitations of current methods

We identified some limitations of current methods, including limitations in identifying smaller and noncontinuous domains, limitations in multislice analysis on large-scale dataset, and gave examples to demonstrate the inadequacies of existing methods on a large-scale dataset.

Limitations in identifying smaller and noncontinuous domains. Current spatial clustering methods incorporating spatial information often results in a preference for continuous results. However, there have been limited investigations into the performance of these methods when applied on data that violate the underlying assumptions of spatial domain continuity. To cover this, we collected an additional 22 data from nonbrain tissues with smaller, noncontinuous tissue domains, including breast cancer[51], liver[52] and pancreatic ductal adenocarcinoma[53] (Methods and Supplementary Table 1). With comprehensive analysis (Supplementary Note 7, Extended Data Fig. 4 and Supplementary Figs. 14–35), we found that all methods encounter challenges when faced with smaller, noncontinuous tissue domains.

Limitations in multislice analysis on large-scale dataset. A growing number of spatial single-cell studies generated spatial data from multiple slices[7,49,54], to construct a large-scale spatial atlas. This research trend necessitates methods that can identify the tissue structures from multiple slices jointly. To cover the multislice and scalability limitations of most current methods, we used a simulated data to show the different types of problems that may occur (Supplementary Note 8 and Supplementary Fig. 36). In addition, we used a recently published large-scale MERFISH dataset containing 378,918 cells from 31 slices, with 374 genes measured[55], to show that all methods encountered time and/or memory issues when applied to the dataset (Supplementary Note 9 and Extended Data Fig. 5).

A divide and conquer strategy enables large-scale scalability

Supplementary Notes 8 and 9 highlighted the inadequacies of existing methods to handle large-scale spatial datasets (Extended Data Fig. 5). Instead of developing new methods to address this, adjusting or combining existing methods may prove to be a more efficient approach. To this end, a divide and conquer strategy was designed to partition the complex task into smaller subtasks that can be solved by existing methods, followed by merging the subresults. A strong base solver is required to solve each subtask. Based on our previous recommendations and data characteristics, SpaceFlow was identified to be the most suitable, given its performance on accuracy, scalability on single-slice small data (especially with imaging-based spatial technologies), and ability to output context-aware representation (Fig. 4). Further details on this divide and conquer strategy can be found in Methods.

https://doi.org/10.1038/s41592-024-02215-8

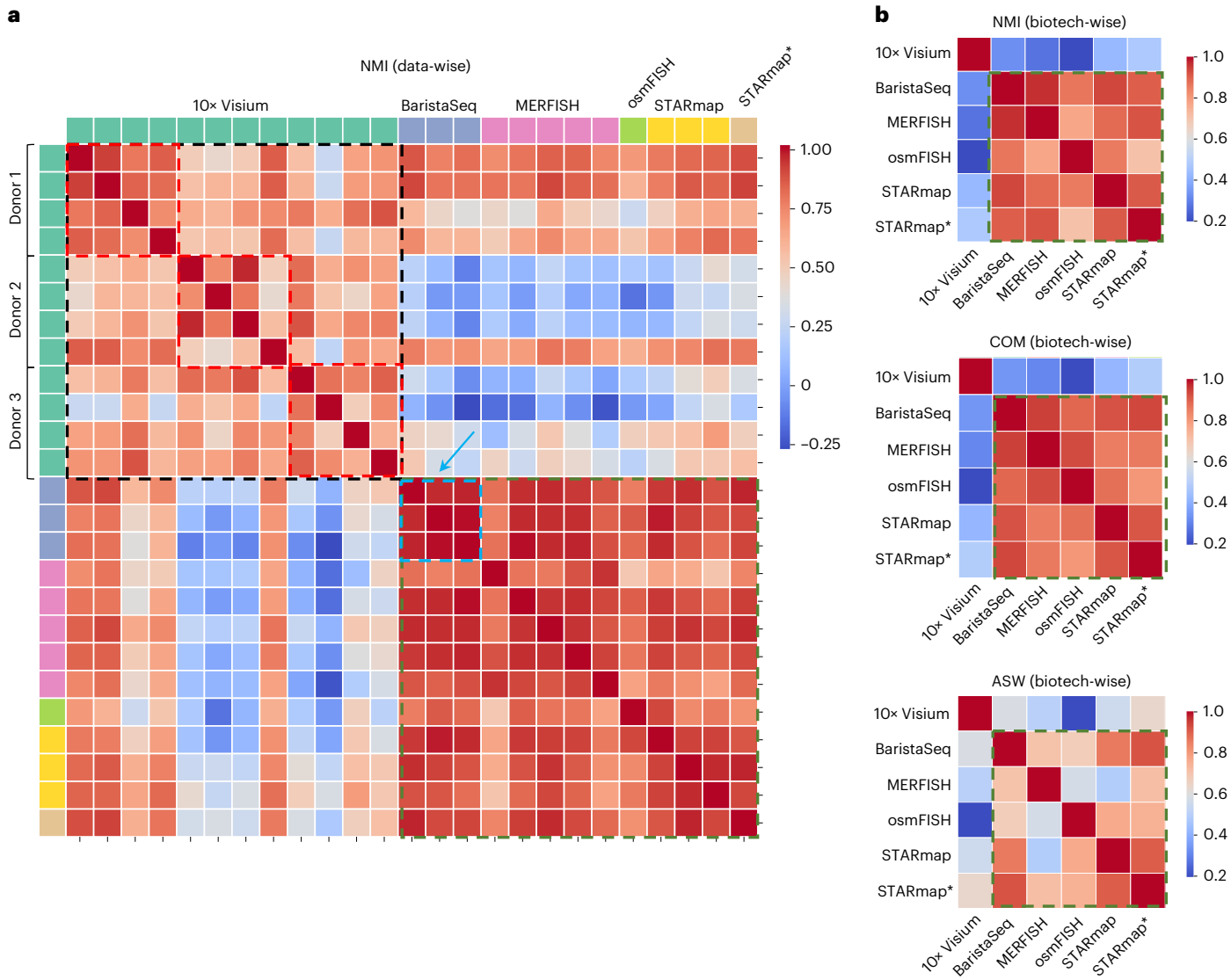


Fig. 3 | Correlations between data regarding methods performance. a, On the heatmap, the entry of row *i* and column *j* is computed by the Spearman's correlation of two vectors. One is formed by the median NMI of all methods applied on data *i*, another is that on data *j* (Methods). The black dashed box was the correlations between data from 10x Visium. The green dashed box was the correlations between data from imaging-based SRT. The blue dashed box pointed by blue arrow was the correlations between data of BaristaSeq. In the 10x Visium dataset, data were collected from three donors, and the red dashed box was the correlations between data from the same donor. **b**, Similar to **a**, but the correlations are between biotechnologies, rather than data. The green dashed box was the correlations between technologies from imaging-based SRT.

The proposed approach, termed SpaceFlow-DC (SpaceFlow Divide and Conquer), was applied to multislice spatial clustering on a large-scale dataset (Extended Data Fig. 5a). Initially, SpaceFlow was applied to each of the 31 slices, respectively, generating a set of context-aware embeddings for each slice (Divide step). As these embeddings were independently trained by graph neural networks, the embeddings and annotation labels were not aligned and exhibited substantial batch effect (Fig. 5a). To address this, Harmony[56] was applied to integrate all the embeddings of the 31 slices (Merge step), producing aligned joint embeddings and spatial clustering results (Fig. 5b). As seen from the cortical layer orders from outermost layer (pia mater) to innermost (corpus callosum), the spatial map of ground truth labels and SpaceFlow-DC-predicted labels exhibited cross-slice-consistent and nicely aligned results (see the consistent color code across slices in Fig. 5c,e). In contrast, the original SpaceFlow exhibited mismatched domain labels across slices (Fig. 5d). This mismatching problem is the common issue when applying single-slice spatial clustering methods on multiple slices.

Quantitative evaluation of the original SpaceFlow and three versions of SpaceFlow-DC (including SpaceFlow-DC-Louvain, SpaceFlow-DC-Leiden and SpaceFlow-DC-mclust; Methods) was conducted to assess spatial clustering accuracy. The approximately equal per-slice accuracy demonstrated that the proposed strategy did not compromise the prediction in every single slice (Fig. 5f). The substantially improved all-slice accuracy indicated that the approach generated accurate and well-aligned spatial domains across all slices (Fig. 5g). The performance of the three versions of SpaceFlow-DC indicated that the clustering methods applied on the context-aware representations did not impact performance and also provided guidance for user's choice. Running times for the three versions of SpaceFlow-DC were recorded as a function of the number of slices (Fig. 5h and Supplementary Fig. 37), indicating that the proposed strategy could handle all cases, with linear time complexity against the data volume, in contrast to all single methods that encountered issues (Extended Data Fig. 5c). Note that all experiments share common computational resources (Methods).

https://doi.org/10.1038/s41592-024-02215-8

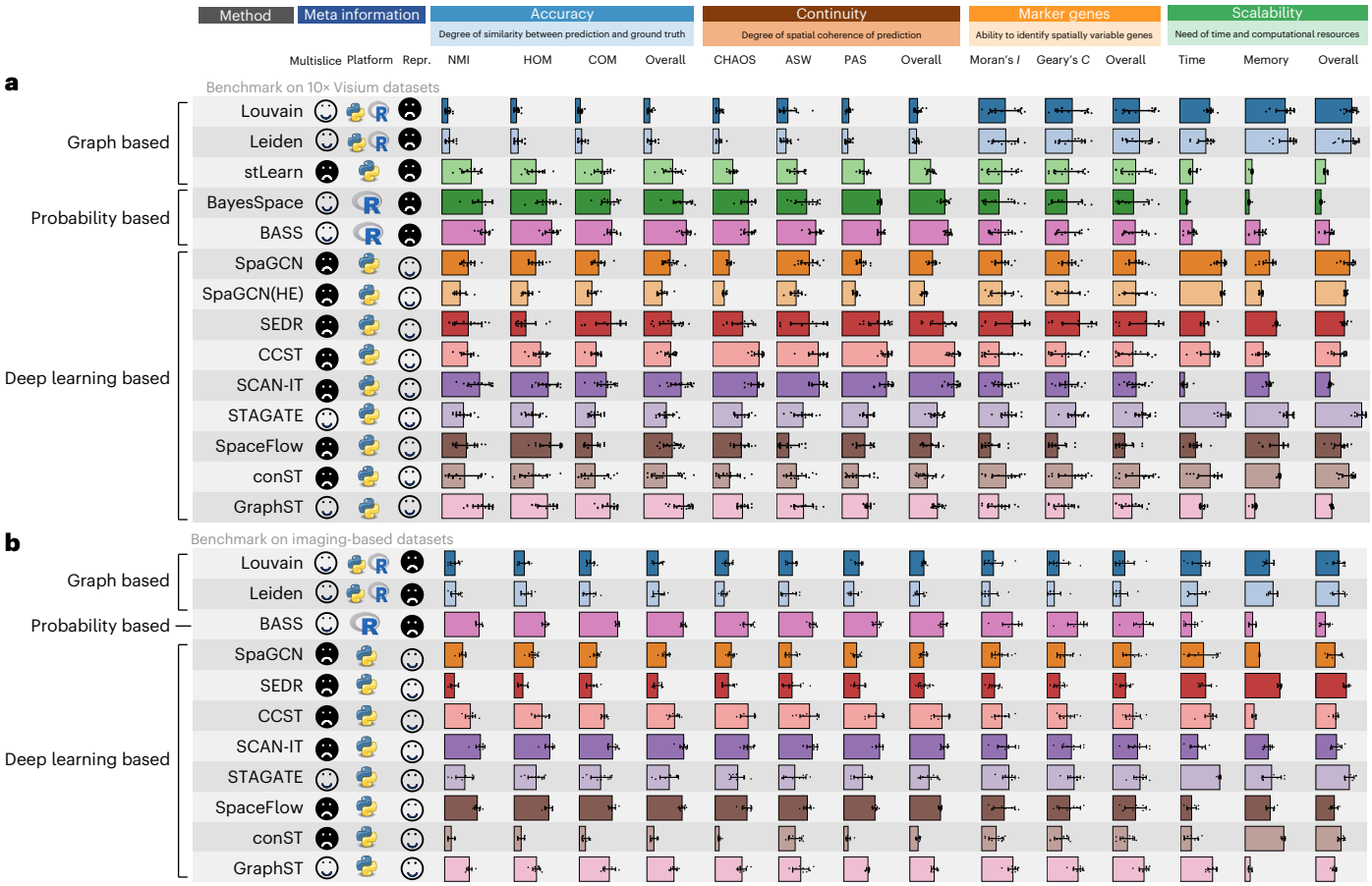


Fig. 4 | Overall performance. a,**b**, The methods information include multislice (whether it supports multislice analysis), platform (Python or R), repr. (whether it can output the context-aware representation), NMI, HOM, COM, CHAOS, ASW, PAS, Moran's *I*, Geary's *C*, running time and peak memory. These metrics are displayed using rank-based scores (Methods). The applied datasets are partitioned into two groups, that is, 10x Visium (**a**) and imaging-based (**b**). Data are presented on the basis of mean and 95% confidence intervals. $N = 12$ data for 10x Visium and $N = 22$ for imaging-based technologies.

Overall, the analysis underscores the usefulness of this benchmark study, demonstrating that the output of this benchmark study can motivate developers to combine existing bioinformatics tools to effectively address the originally impossible mission. The proposed divide and conquer strategy provides a strong solution for the ever-increasing amount of large-scale datasets from a bunch of slices (Extended Data Fig. 5b).

Impact of pre- and postprocessing and robustness

We evaluated the impact of different preprocessing and postprocessing approaches and robustness against different data characteristics (Fig. 6).

In the context of spatial clustering with spatial transcriptomics data, various preprocessing and postprocessing methods may impact the prediction results. However, current methods use inconsistent pre- and postprocessing steps, leading to unclear and uncontrolled performance. For preprocessing, we considered highly variable gene (HVG) selection, spatially variable genes (SVG) selection and no selection. For postprocessing, we considered refinement (K-nearest neighbors (KNN)-based method to refine the identified spatial domain labels to smooth them in space, proposed by SpaGCN[12]), and no refinement. These inconsistent pre- and postprocessing steps outside the main body of spatial clustering methods motivated us to assess the effects of pre- and postprocessing approaches on the performance for different methods across datasets. Through a comprehensive analysis (Fig. 6a–d, Supplementary Note 10 and Supplementary Figs. 38–41) of the preprocessing steps, we found that approximately half of the tested methods achieved better performance without any gene selection. While regarding the postprocessing step, all methods benefit from a post hoc spatial smoothing step.

One can also encounter a variety of spatial transcriptomics data with different characteristics in their studies, therefore simulating new data with variable parameters might be useful for a wider reference. Considering three important properties of the spatial transcriptomics dataset, we simulated data with different gene expression matrix sparsity, spatial resolutions and gene numbers (Methods). In addition, we also tested the robustness of spatial clustering methods against adding noise (Methods). We analyzed the robustness of different methods against these factors (Fig. 6e, Supplementary Note 11 and Supplementary Figs. 42–51). Regarding the impact of gene expression matrix sparsity, for most methods, the performance reduced with increasing expression matrix sparsity, except CCST, Louvain and Leiden. Regarding the impact of spatial resolution, most methods' performance slowly increased as the resolution became lower, except Louvain, Leiden, SCAN-IT and SpaceFlow. Regarding the impact of gene subsampling, most methods displayed higher performance with larger number of genes. Regarding the impact of adding noise, we identified a consistent robustness across spatial clustering methods.

Discussion

In this study, we conducted a benchmark analysis of 13 methods for identifying spatial domains using spatial transcriptomics data from various spatial technologies. We evaluated these methods on the basis of accuracy, continuity, marker score and scalability, providing a comprehensive framework for spatial clustering evaluation that can assist researchers

https://doi.org/10.1038/s41592-024-02215-8

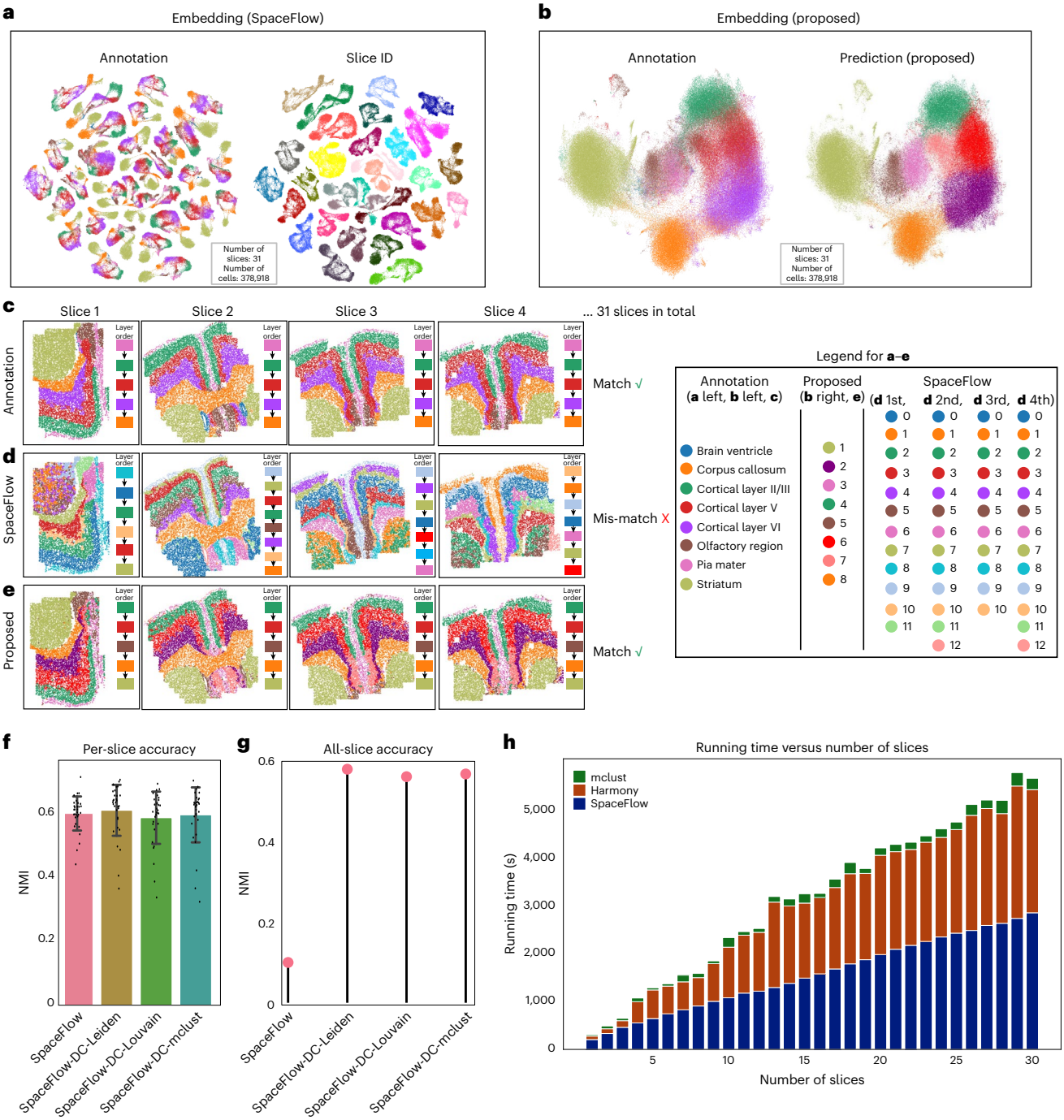


Fig. 5 | The large-scale datasets are solved by proposed approach. The dataset used in this figure is from Extended Data Fig. 5a. **a**, The visualization is generated by applying UMAP on the SpaceFlow embedding of all slices, colored by ground truth annotation (left) and slice ID (right). **b**, The visualization is generated by applying UMAP on the SpaceFlow-DC embedding (proposed) of all slices, colored by ground truth annotation (left) and domain predicted by SpaceFlow-DC (right). **c**–**e**,: spatial plots of 4 of 31 slices showed that the expert-annotated ground truth (**c**) and SpaceFlow-DC domain labels (**e**) are matched across slices, but the original SpaceFlow cannot generated aligned labels across slices (**d**). From the outermost layer to the innermost layer, the labels (indicated by different colors) order also showed that original SpaceFlow (**d**) cannot generate aligned labels across slices. **f**, For each slice independently, NMI was recorded on SpaceFlow and three versions of SpaceFlow-DC (proposed). Data are presented on the basis of mean and 95% confidence intervals. $N = 31$ slices. **g**, For all slices, NMI was recorder on SpaceFlow and three versions of SpaceFlow-DC (proposed). **h**, Running time of three different steps (that is, SpaceFlow, Harmony and mclust) of SpaceFlow-DC was recorded for different subsets of slices of whole dataset.

in selecting the optimal spatial clustering tool for their spatial transcriptomics data. Our results indicate that no single method is universally effective across all datasets, and the optimal method is dependent on data characteristics. Our comparative analysis identified several limitations of current methods, including the challenges in identifying small regions, the lack of multislice analysis capability and large-scale scalability issues. With increasing requirements for memory and computation time of spatial transcriptomics data, greater scalability is desirable to meet these demands. Regarding some identified limitations, we proposed a 'divide and conquer' approach to demonstrate that our benchmark study

https://doi.org/10.1038/s41592-024-02215-8

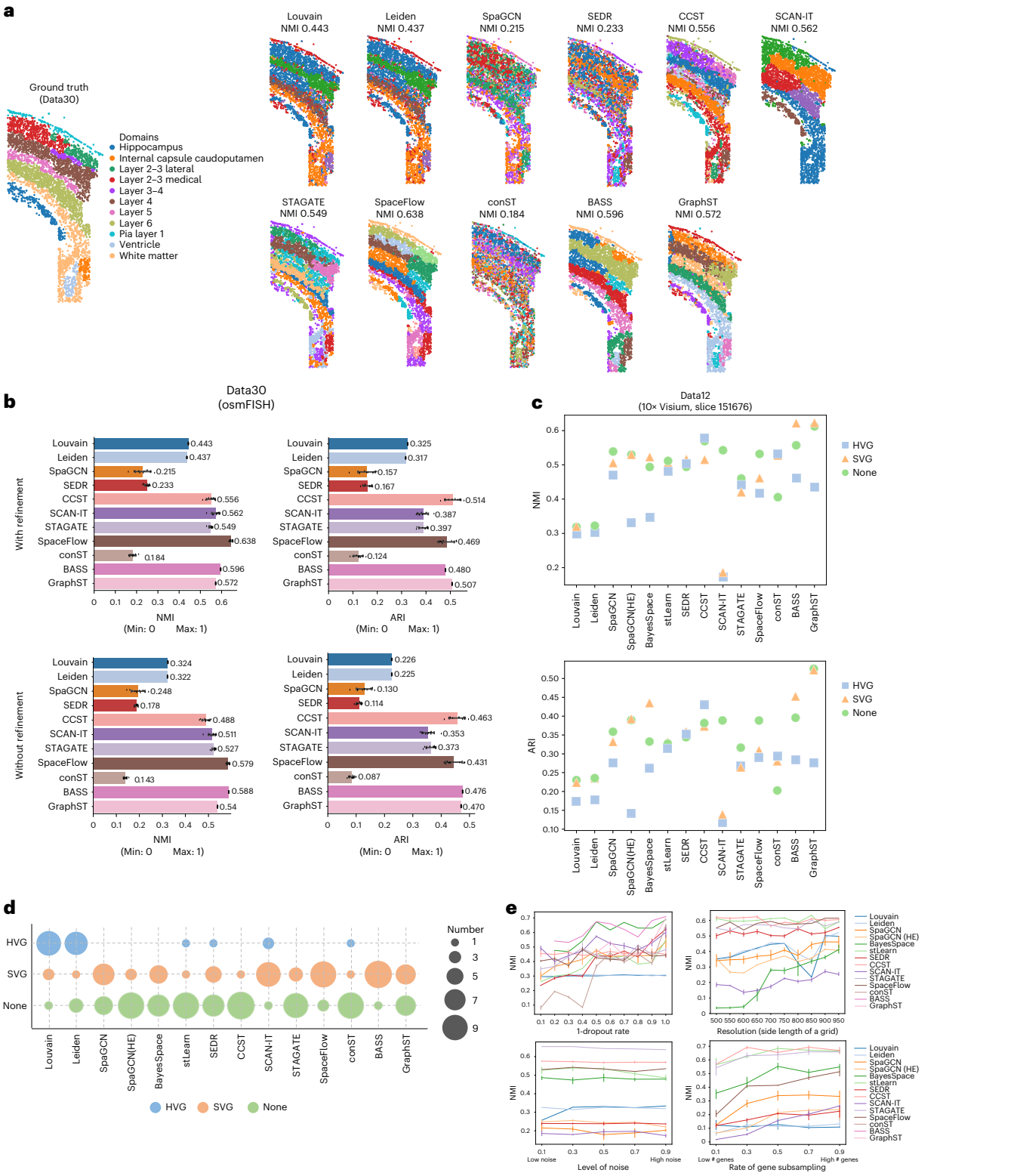


Fig. 6 | The impact of pre- and postprocessing. a, The ground truth and methods' output on osmFISH data. Note that all results are postprocessed by the KNN-based refinement approach. **b:** Quantitative evaluations of different methods with (top) and without (bottom) refinement were recorded on osmFISH data. Data are presented on the basis of mean and 95% confidence intervals. $N = 10$ independent runs. **c**, Quantitative evaluations were recorded on 10x Visium data to explore the impact of different preprocessing approaches, that is, HVG selection, SVG selection and no preprocessing (None). **d**, Methods performance with different preprocessing approaches. The bubble chart exhibits how various combinations of preprocessing approaches and spatial clustering methods perform. The size of the dots represents the quantity of datasets in which this preprocessing method yielded greater NMI values compared to the other two preprocessing methods. **e**, Evaluation of methods robustness against four different factors, that is, expression matrix sparsity (top left), spatial resolution (top right), level of noise (bottom left) and gene subsampling rate (bottom right). For each panel, the *x* axis represents level of factors, and the *y* axis represents the NMI values. Different colors represent different methods. Data are presented on the basis of mean and 95% confidence intervals. $N = 10$ independent runs.

can motivate method bioinformaticians to combine existing tools to effectively address the challenges posed by large-scale spatial datasets.

We also focused on the incorporation of additional staining images such as hematoxylin and eosin (H&E). In line with ensuring a fair comparison, we included both versions of SpaGCN—with and without H&E staining—in our benchmarking process. Interestingly, our analysis, corroborated by findings from other studies, indicates that the addition of H&E staining images does not consistently enhance the performance of SpaGCN. It appears that H&E images might contain certain information that does not directly aid in the improved identification of tissue structures. These data may inadvertently introduce extraneous noise, thus influencing the overall performance. Another possible explanation for these findings could be related to the way how H&E images were modeled. There might be room for further refinement to optimally exploit the additional data that H&E staining provides.

There are some limitations of this study. First, the fast-paced nature of advancements in the spatial transcriptomics field and the limited ground truth annotations have led to the exclusion of numerous emerging data types from our study. Second, our study predominantly concentrated on spatial clustering methods applied to spatially resolved transcriptomics data. This focus inadvertently neglected the incorporation of other cutting-edge spatial data, such as spatial proteomics and spatial multiomics data, which could potentially provide complementary insights. To address these limitations, future studies should actively strive to encompass more extensive array of spatial data types and spatial computational methods. Given the similar data formats across various spatial data types, we anticipate that upcoming studies will evaluate more comprehensive computational methods across a wider spectrum of spatial data.

This study can also inspire thoughts for designing future benchmark studies, and the following points might be considered. Regarding metrics, the key measure should be the similarity between the method's prediction and the ground truth. In our manuscript, for example, we used NMI, HOM and COM to evaluate this similarity. If designing a benchmark for detecting SVGs, for example, metrics such as precision, recall or *F*-measure would be appropriate as they evaluate the agreement between the detected gene set and the truly SVG set. Additional metrics that reflect some properties of the prediction, independent of ground truth, can also be helpful. In our manuscript, metrics like CHAOS, PAS and ASW measure the spatial continuity of the predicted domains. This principle of high spatial continuity is inherent in tissue biology; however, it may not always indicate a good prediction. For example, assigning an identical label to all cells would result in the highest spatial continuity. Hence, such metrics should be used to supplement the primary similarity-based metrics. For SVG benchmarking, classification accuracy (to predict spatial domains) using detected SVGs as features is another side measure of quality SVG prediction[12], but remember that some SVGs may not correspond to spatial domains. Scalability measures such as running time and peak memory usage are also important. These metrics are frequently used in benchmark studies across various computational tasks, including single-cell data integration[57], spatial transcriptomics data imputation or deconvolution[42], and trajectory inference[58]. Regarding dataset selection, the availability of reliable 'ground truth' is the first consideration since performance evaluation depends on comparing method predictions with this ground truth. This was our primary criterion in our study. However, for tasks like SVG identification that lack real datasets with well-established ground truth, simulated data can be used as a substitute until real datasets with reliable ground truth become available. The diversity of datasets should also be considered. Given the varying data-generation protocols across biotechnologies and the ever-evolving landscape of these technologies[43], it is advantageous to cover a broad range of data types in the benchmark study. This was another major consideration in our study design. Lastly, the choice of metrics and datasets should consider the extent of their recognition in the field. While not all selected metrics or datasets need to be widely used, a substantial portion should be familiar to the research community. This is beneficial when comparing the new benchmark study with existing benchmarks presented in published method papers.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02215-8.

References

1. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-023-00580-2 (2023).
2. Seferbekova, Z., Lomakin, A., Yates, L. R. & Gerstung, M. Spatial biology of cancer evolution. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-022-00553-x (2022).
3. Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-022-00515-3 (2022).
4. Zeng, H. et al. Spatially resolved single-cell translatomics at molecular resolution. *Science* **380**, eadd3067 (2023).
5. Shi, H. et al. Spatial atlas of the mouse central nervous system at molecular resolution. *Nature* https://doi.org/10.1038/s41586-023-06569-5 (2023).
6. Chen, A. et al. Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex. *Cell* **186**, 3726–3743 e3724 (2023).
7. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).
8. Zhang, M. et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).
9. Chang, Y. et al. Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning. *Comput. Struct. Biotechnol. J.* **20**, 4600–4617 (2022).
10. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with adaptive graph attention auto-encoder. *Nat. Commun.* https://doi.org/10.1038/s41467-022-29439-6 (2021).
11. Fu, H. et al. Unsupervised spatial embedded deep representation of spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.15.448542 (2021).
12. Hu, J. et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
13. Li, J., Chen, S., Pan, X., Yuan, Y. & Shen, H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Nat. Comput. Sci.* **2**, 399–408 (2022).
14. Yuan, Z. et al. SOTIP is a versatile method for microenvironment modeling with spatial omics data. *Nat. Commun.* **13**, 7330 (2022).
15. Yang, M. et al. Position-informed contrastive learning for spatially resolved omics deciphers hierarchical tissue structure at both cellular and niche levels. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-1067780/v1 (2022).
16. Cable, D. M. et al. Cell type-specific inference of differential expression in spatial transcriptomics. *Nat. Methods* **19**, 1076–1087 (2022).
17. Zeng, H. et al. Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease. *Nat. Neurosci.* https://doi.org/10.1038/s41593-022-01251-x (2023).
18. Palla, G., Fischer, D. S., Regev, A. & Theis, F. J. Spatial components of molecular tissue biology. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01182-1 (2022).

19. Rao, A., Barkley, D., Franca, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
20. Cheng, A., Hu, G. & Li, W. V. Benchmarking cell-type clustering methods for spatially resolved transcriptomics data. *Brief. Bioinform.* **24**, bbac475 (2023).
21. Xu, Z. et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res.* **52**, D1053–D1061 (2024).
22. Long, B., Miller, J. & The SpaceTx Consortium. SpaceTx: a roadmap for benchmarking spatial transcriptomics exploration of the brain. Preprint at https://arxiv.org/abs/2301.08436 (2023).
23. Megill, C. et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.05.438318 (2021).
24. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz934 (2019).
25. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
26. Yuan, Z. et al. SODB facilitates comprehensive exploration of spatial omics data. *Nat. Methods* https://doi.org/10.1038/s41592-023-01773-7 (2023).
27. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792 e1721 (2022).
28. Chen, X., Sun, Y.-C., Church, G. M., Lee, J. H. & Zador, A. M. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* **46**, e22 (2018).
29. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. Y. & Zhuang, X. W. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* https://doi.org/10.1126/science.aaa6090 (2015).
30. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
31. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
32. Ren, H., Walker, B. L., Cang, Z. & Nie, Q. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat. Commun.* **13**, 4076 (2022).
33. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
34. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
35. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00935-2 (2021).
36. Pham, D. et al. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat. Commun.* **14**, 7739 (2023).
37. Zixuan, C., Ning., X., Nie, A., Xu, M. & Zhang, J. SCAN-IT: domain segmentation of spatial transcriptomics images by graph neural network. In *32nd British Machine Vision Conference* https://www.bmvc2021-virtualconference.com/conference/papers/paper_1139.html (2021).
38. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).
39. Zong, Y. et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.01.14.476408 (2022).
40. Li, Z. & Zhou, X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biol.* **23**, 168 (2022).
41. Long, Y. et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat. Commun.* **14**, 1155 (2023).
42. Li, B. et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19**, 662–670 (2022).
43. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
44. Rosenberg, A. & Hirschberg, J. V-measure: a conditional entropy-based external cluster evaluation measure. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* 410–420 (2007).
45. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat. Commun.* **13**, 7203 (2022).
46. Zuo, C. et al. Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nat. Commun.* **13**, 5962 (2022).
47. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
48. Geary, R. C. The contiguity ratio and statistical mapping. *Incorp. Stat.* **5**, 115–146 (1954).
49. Fang, R. et al. Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science* **377**, 56–62 (2022).
50. Moffitt, J. R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
51. Andersson, A. et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* **12**, 6012 (2021).
52. Guilliams, M. et al. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* **185**, 379–396. e338 (2022).
53. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
54. Lohoff, T. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01006-2 (2021).
55. Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* https://doi.org/10.1016/j.cell.2022.12.010 (2022).
56. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
57. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
58. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

Methods

Settings of benchmarked methods

Each evaluation in our study involved ten replicated runs of each method. To show our general criteria for parameter choice, we have detailed our process in three categories:

Parameters determining number of spatial domains. For the parameters that directly influence the number of spatial domains in our clustering results, we adopted two approaches based on the capabilities of the methods. (1) For algorithms where we could directly input the expected number of spatial domains (for example, SpaGCN and BayesSpace), we set the parameter to the actual number of domains obtained from our ground truth. (2) For those algorithms that could only accept clustering resolution, we searched the resolution that best matches the expected number of spatial domains.

Data-related parameters. This category is related to parameters influenced by the spatial organization and feature dimensions of different datasets. For these parameters, we adopted a parameter search strategy. Such parameters are involved in almost every method, such as the number of neighbors when constructing the spatial graph, and the number of principal components (n_pcs) when processing the feature space. When setting the search ranges, we consider covering all the default values of all methods. For example, across all methods that have n_pcs as parameter, the max default value of n_pcs is 200 (CCST), and the min default value of n_pcs is 15 (BayesSpace), so we set the searching range of n_pcs to 15–200 (step 5).

Network-related parameters. This last category of parameters includes those controlling the network layers, the number of neurons in hidden layers, and the training stopping criteria. For these parameters, we adhered to the original authors' recommended settings.

In the following, we explained the basic principles of different methods. Full parameter descriptions, choice principle and search ranges are available at Supplementary Table 3.

Louvain. Louvain[33] is a nonspatial clustering algorithm that allocates each spot to a distinct community and optimizes modularity by iteratively merging and splitting communities until the desired clustering results are achieved. We followed the guidelines outlined on the SCANPY website[59].

Leiden. Leiden[33,34] is another nonspatial clustering algorithm that is similar to Louvain but can also merge similar communities to enhance clustering results further. We followed the guidelines outlined on the SCANPY website[60]. The relationships between Louvain and Leiden are explained in Supplementary Note 12.

SpaGCN. SpaGCN[12] is a spatial clustering method that utilizes graph convolutional networks (GCNs) to integrate gene expression, spatial location and histological information to cluster spots into different spatial domains using unsupervised iterative clustering. This version of SpaGCN is for data without H&E images.

SpaGCN(HE). SpaGCN(HE)[12] is a spatial clustering method that employs GCNs to integrate gene expression, spatial location and histological information for unsupervised iterative clustering. This method utilizes an undirected weighted graph to determine the Euclidean distance between spots in the graph on the basis of their spatial coordinates (x, y) and the three-dimensional coordinate z, obtained from the RGB values in the histological image. This version of SpaGCN is for data with H&E images.

BayesSpace. BayesSpace[35] is a full Bayesian statistical method that incorporates a low-dimensional representation of the gene expression matrix to model spatial clustering, encouraging neighboring points to belong to the same cluster through a spatial prior. This method requires spatial data in the form of spot locations, so it is not suitable for data that do not use 'spot'.

StLearn. stLearn[36] utilizes spatial morphological gene expression (SME)-normalized data to perform unsupervised clustering by grouping similar points into clusters. It also utilizes the spatial information of these clusters to identify subclasses within the organization. This method requires histological information (for example, H&E images) as input and thus cannot be used on data lacking such information.

SEDR. SEDR[11] is an unsupervised spatial clustering algorithm that utilizes a deep autoencoder to construct a low-dimensional latent representation of gene expression data. This representation is then integrated with the corresponding spatial information using a variational graph autoencoder, allowing for simultaneous spatial embedding during the clustering process.

CCST. CCST[13] is a spatial clustering method that encodes cell location and gene expression information into an adjacency matrix and gene expression matrix. These two matrices are then input into a Deep Graph Infomax network to obtain cell node embeddings containing spatial and gene expression information. The embeddings are then downscaled by principal component analysis (PCA) and clustered using the k-means++ algorithm.

SCAN-IT. SCAN-IT[37] is a method that utilizes an image segmentation approach to solve the spatial clustering problem. It treats cells as pixels in an image and represents gene expressions within cells as different channels similar to RGB channels. It constructs a geometry-aware spatial proximity graph using alpha complex, generates low-dimensional embeddings of the spots through the application of Deep Graph Infomax, and uses common clustering algorithms to obtain clustering results based on the resulting low-dimensional representations. The step with the SOMDE[61] algorithm prolongs the overall actual running time of SCAN-IT.

STAGATE. STAGATE[38] is a spatial clustering method that transforms spatial location information into a spatial neighbor network and utilizes gene expression information and the network to train a graph attentional self-coding neural network. This network generates low-dimensional latent embeddings that integrate spatial and expression information. These embeddings can then be used for clustering.

SpaceFlow. The SpaceFlow[32] method utilizes deep GNNs to merge gene expression similarity and spatial information, generating spatially consistent low-dimensional embeddings through spatial regularization. These embeddings can then be used for clustering.

conST. conST[39] is a spatial clustering method that uses contrastive learning to integrate multiple modalities of SRT data, including gene expression, spatial information and morphology. The method applies data augmentation and learns low-dimensional embeddings by minimizing or maximizing the mutual information between different embeddings using three levels of comparison learning. These embeddings can then be used for clustering.

BASS. BASS[40] is a spatial clustering method that facilitates multiscale and multisample analysis of spatial transcriptomics data. It utilizes a Bayesian hierarchical modeling framework to perform clustering analysis.

GraphST. GraphST[41] is a method that utilizes a graph self-supervised comparative learning model for spatial clustering. It integrates GNNs and self-supervised comparative learning to effectively learn spot

representations in spatial transcriptomics data by modeling both gene expression and spatial localization information. These embeddings can then be used for clustering.

Impact of preprocessing approaches

We employed the 10x Visium dataset to investigate the impact of three gene selection methods, namely HVGs, SVGs and no processing (None). For HVG, we utilized the 'highly_variable_genes' function from the SCANPY package to identify the top 3,000 HVGs. For SVGs, we used the SPARK[62] algorithm[63] to extract the first 3,000 SVGs.

Impact of postprocessing approaches

We used the KNN-based refinement approach as implemented SpaGCN package to postprocess all the results of identified spatial domain across all datasets[12]. We then compared the effects of applying the postprocessing method versus not applying it. To achieve this, we calculated the adjacency matrix of the original spatial transcriptomics data and applied the spg.refine() function available in SpaGCN package for postprocessing. We then explored the impact of this postprocessing on our results.

Real datasets

Across the manuscript, when referring a spatial transcriptomics data of a single slice, we call it 'data', and when referring a collection of data from the same publication and sharing the same technology, we call it 'dataset'. The real data are sequentially labeled, ranging from Data1 to Data56. These data also have their original names in their respective source publications. A map pertaining to these original names can be found in Supplementary Table 1. Note that Data35 to Data56 were used for testing methods when the spatial continuity assumption was violated. Data57 to Data87 were used for testing methods on large-scale multislice dataset.

Data1 to Data12. The 10x Visium dataset (also known as SpatialLIBD or DLPFC) is the most widely used benchmark dataset among spatial clustering methods[25]. This dataset contains 12 data of human postmortem DLPFC tissue sections, from three independent neurotypical adult donors, all profiled using 10x Visium with paired H&E images. Like other 10x Visium datasets, each measurement unit is a spot. The number of barcoded array spots for the 12 samples are 4,226 (Data1), 4,384 (Data2), 4,789 (Data3), 4,634 (Data4), 3,661 (Data5), 3,498 (Data6), 4,110 (Data7), 4,015 (Data8), 3,639 (Data9), 3,673 (Data10), 3,592 (Data11) and 3,460 (Data12). The number of unique genes is 33,538. This dataset was downloaded from ref. 64, containing H&E images of different resolutions for each data, as well as spot-wise region annotation. The region annotation ranges from layer 1 to layer 6, and white matter.

Data13 to Data21. The Stereo-Seq dataset is obtained by high-resolution full-transcriptome coverage technologies (Stereo-Seq technology)[27]. This dataset contains nine data obtained from different samples of mouse embryo. The numbers of spots for the nine data are 5,913 (Data13), 5,292 (Data14), 4,356 (Data15), 5,059 (Data16), 5,797 (Data17), 18,408 (Data18), 18,647 (Data19), 18,670 (Data20) and 8,494 (Data21). The number of genes for the nine samples are 25,568 (Data13), 23,756 (Data14), 24,107 (Data15), 24,238 (Data16), 23,398 (Data17), 25,201 (Data18), 25,544 (Data19), 25,647 (Data20) and 22,385 (Data21). This dataset was downloaded from ref. 65, along with the region annotations.

Data22 to Data24. The BaristaSeq dataset is an imaging-based spatial transcriptomics dataset obtained by BaristaSeq technology[28]. This dataset contains three data from mouse primary cortex. The numbers of cells for the three samples are 1,525 (sample Slice_1), 2,042 (sample Slice_2) and 1,690 (sample Slice_3). The number of unique genes is 79. This dataset was downloaded from ref. 66 (ref. 22).

Data25 to Data29. The MERFISH dataset is an imaging-based spatial transcriptomics dataset published in 2018[50]. Among all slices, five slices were annotated with region labels[40]. The numbers of cells are 5,488 (Data25), 5,557 (Data26), 5,926 (Data27), 5,803 (Data28) and 5,543 (Data29). The number of unique genes is 155. This dataset was downloaded from ref. 67.

Data30. The osmFISH dataset is an imaging-based spatial transcriptomics dataset published in 2018 (ref. 30). This dataset contains one data from mouse somatosensory cortex, published with the paper presenting osmFISH technology. The number of cells is 4,839. The number of unique genes is 33, selected from published single-cell RNA-sequencing datasets of mouse somatosensory cortex. We downloaded the dataset from ref. 68.

Data31 to Data33. STARmap is an imaging-based spatial transcriptomics dataset published with the paper presenting STARmap technology[31]. This dataset contains three data from mouse medial prefrontal cortex, with expert annotations of layers[40]. The numbers of cells are 1,049 (Data31), 1,053 (Data32) and 1,088 (Data33). The number of unique genes is 166. This dataset was downloaded from ref. 69.

Data34. STARmap* is a version of the STARmap measuring 1,020 genes. This dataset contains one data from mouse visual cortex with layer annotations, which was downloaded from ref. 69.

Data35 to Data41. This dataset is HER2-positive breast tumors using spatial transcriptomics technology. This dataset contains seven data. The number of spots for the seven samples is 341 (Data35), 269 (Data36), 167 (Data37), 255 (Data38), 534 (Data39), 659 (Data40) and 530 (Data41). The number of genes for the seven samples is 15,045 (Data35), 15,109 (Data35), 15,557 (Data35), 15,661 (Data35), 15,701 (Data35), 14,861 (Data35) and 15,029 (Data35). This dataset was downloaded from ref. 70 (ref. 51).

Data42 to Data54. This dataset is from liver tissues using 10x Visium technology. This dataset contains 13 data, 8 of which are mouse liver (Data42 to Data49) and 5 are human liver (Data50 to Data54). The number of spots for the 13 data is 1,293 (Data42), 1,363 (Data43), 1,316 (Data44), 1,790 (Data45), 1,121 (Data46), 2,002 (Data47), 1,780 (Data48), 1,768 (Data49), 626 (Data50), 371 (Data51), 1,759 (Data52), 658 (Data53) and 919 (Data54). The number of unique genes for mouse samples is 31,053, and the number of unique genes for human samples is 32,738. This dataset was downloaded from https://www.livercellatlas.org/ (ref. 52).

Data55 to Data56. This dataset is from human pancreatic ductal adenocarcinomas using spatial transcriptomics technology. This dataset contains two data. The number of spots for the two data is 428 (Data55) and 224 (Data56). The number of unique genes is 19,738. The data are available at GSE111672 (ref. 53).

Data57 to Data87. This dataset is from mouse aging MERFISH dataset containing 31 data. This dataset is for testing the current methods on large-scale dataset containing a high number of slices. The information is available at Extended Data Fig. 5a. The dataset was downloaded from ref. 71 (ref. 55).

Simulated data

Varying spatial resolutions. We used osmFISH dataset[30] to assess the influence of different spatial resolutions on each method's performance. To exclusively examine the impact of spatial resolution, we maintained other data parameters constant. Our approach involved defining a square grid with n-um windows (n = 50, 55, 60, 65, 70, 75, 80, 85, 90, 95) to partition the osmFISH data. To prevent extremely low

gene expression values in certain spots, we excluded spots containing one cell or fewer, resulting in multiple cells simulating each spot. The position of each spot was determined as the center of its corresponding grid, and we simulated the spot's counts by summing the expression values of all cells within the grid. The ground truth for each spot was determined using the majority voting approach.

Varying expression matrix sparsity. To investigate the impact of gene expression matrix sparsity, we assessed the performance of each method using the 10x Visium dataset. We randomly removed count values at different rates (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9). For instance, when the rate is set to 0.9, it implies that each value in the gene expression matrix has a 0.9 probability of being set to 0. Conversely, a rate of 0 means that the gene expression matrix remains unchanged.

Varying gene subsampling rate. We performed gene subsampling on data acquired from diverse spatial transcriptomics technologies (10x Visium, BaristaSeq, MERFISH, STARmap and STARmap*). Our analysis excluded the osmFISH dataset, as it only contained 33 genes. To generate the different gene subsampled datasets, we applied a range of subsampling rates (10%, 30%, 50%, 70% and 90%). The desired number of genes for each dataset was obtained by multiplying the total number of genes in the original data by the subsampling rate. Then, we randomly selected (without replacement) genes from the dataset on the basis of the calculated gene count.

Adding various levels of noise. We explore the impact of different levels of noise on the performance for each method. We performed on osmFISH dataset. To generate varying levels of noise (10%, 30%, 50%, 70% and 90%), we utilized a Poisson distribution with the mean parameter set as the product of the noise level and the mean gene expression level.

Benchmark metrics

In our study, we used the following metrics to assess each method.

NMI. NMI quantifies the similarity between two clusterings, and ranges from 0 to 1. The closer the NMI is to 1, the better agreement between two clustering assignments. NMI has been widely used in single-cell clustering accuracy evaluation[72–74] and was also used to assess the performance of spatial domain identification algorithms[46]. We use NMI to assess the agreement between the ground truth domain label and tissue structure identification result computed by each method. Suppose P is the spatial domain clustering result, T is the ground truth clustering label, their entropies are $H(P)$ and $H(T)$, respectively, and the mutual information is $\mathrm{MI}(P, T)$, then NMI is computed as

$$\mathrm{NMI} = \frac{\mathrm{MI}(P, T)}{\sqrt{H(P)H(T)}}.$$

ARI. The adjusted Rand index (ARI) is a measure of similarity between two clusterings that takes into account the possibility of chance agreement between the two clusters. It is commonly used to evaluate the performance of clustering algorithms. ARI ranges from 0 to 1, where ARI values closer to 1 indicate more similar clustering results. To calculate the ARI, the contingency table is constructed by comparing the true domain labels with the predicted tissue structure identification result for each spot. The contingency table contains four entries: TF is the number of spots that are in the same cluster in both the true and predicted clustering, TN is the number of pairs that are in different clusters in both the true and predicted clustering, FN is the number of pairs that are in the same cluster in the true clustering but in different clusters in the predicted clustering, and FP is the number of pairs that are in different clusters in the true clustering but in the same cluster in the predicted clustering. ARI is computed as

$$\mathrm{ARI} = \frac{\mathrm{TP} + \mathrm{TN} - E}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} - E}$$

where E is the expected value of the index, which is the value that would be obtained if the clustering were completely random. The expected value of the index is calculated as

$$E = \frac{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN}) \times (\mathrm{FP} + \mathrm{TN})}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

HOM. The HOM score is a metric that quantifies the homogeneity of a cluster labeling when compared to a known ground truth. A clustering outcome is considered homogeneous if all of its clusters exclusively comprise data points belonging to a single class. The HOM score is expressed as a value between 0 and 1, with 1 representing a perfectly homogeneous labeling. We followed the instructions at ref. 75.

COM. The COM score is a metric that evaluates the completeness of a cluster labeling with respect to a ground truth. A clustering result is considered complete if all data points that belong to a certain class are grouped into the same cluster. The COM score ranges from 0 to 1, with a value of 1 indicating a perfect and complete labeling. We followed the instructions at ref. 76.

CHAOS. The CHAOS score is a metric that has been used to assess the performance of spatial continuity in mass spectrometry imaging field[77,78], and spatial transcriptomics field[45]. A lower CHAOS indicates a better continuity result of spatial domain identification. To apply CHAOS to quantify each spatial clustering method's performance, we first build a 1-nearest neighbor (NN) graph for each data. Specifically, each cell is connected to another cell, which has minimum Euclidean distance in physical space. With this 1-NN graph, suppose d_{ij} is the Euclidean distance between cell_i and cell_j in physical space, we compute w as

$$w_{kij} = \begin{cases} d_{ij}, & \text{if } \mathrm{cell}_i \text{ and } \mathrm{cell}_j \text{ are connected in the 1NN graph in cluster } k \\ 0, & \text{otherwise} \end{cases}$$

Suppose n_k is the number of cells in the k th spatial domain, n is the total number of cells in the data, and K is the number of unique spatial domains. CHAOS is computed as

$$\mathrm{CHAOS} = \frac{\sum_{k=1}^{K} \sum_{ij}^{n_k} w_{kij}}{N}.$$

PAS. The PAS score has been used in quantifying the spatial homogeneity of spatial domain identification algorithms in the field of spatial transcriptomics[45]. A lower PAS score indicates a better continuity of detected spatial domains, which expects higher cell homogeneity within spatial domains. The PAS score is calculated as the percentage of cells with a spatial domain label that is different from at least six of its neighboring ten cells.

ASW. ASW[79] is originally used for evaluating the degree of agreement between clustering labels and an embedding (or distance matrix). We extend ASW to evaluate the spatial coherence of predicted domains regarding to the physical space. The value of ASW ranges from −1 to 1 (we rescaled ASW to 0–1), and the closer ASW is to 1, the better the performance. To define ASW, silhouette width (SW) should be defined first, and ASW could be computed by averaging SWs across all cells. Suppose a is the mean distance between a cell and all other cells in the same spatial domain, and b is the mean distance between a cell and all other cells in the next nearest cluster, then SW of a cell is computed as

$$\mathrm{SW} = \frac{b - a}{\max(a, b)}.$$

Moran's *I*. Moran's *I* is a metric to quantify the degree of spatial autocorrelation in spatial statistics, and has been widely used to evaluate whether the detected SVGs exhibit an organized spatial expression pattern in the spatial omics field[12,80]. The value of Moran's *I* ranges from −1 to 1, and a value close to 1 indicates a clear spatial gene expression pattern, a value close to 0 indicates a random spatial gene expression pattern, and a value close to −1 indicates a gene expression pattern that looks like a chess board. In our case, Moran's *I* will be used to evaluate the spatial autocorrelation for each SVG computed by each method. For one gene, suppose x_i and x_j are the gene expression values of cell_i and cell_j, $\bar{x}$ is the average gene expression value of the gene, and N is the number of cells. Then Moran's *I* is computed as

$$\text{Moran's } I = \frac{N}{W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

where

$$w_{ij} = \begin{cases} 1, \& \text{ if } i \text{ and } j \text{ are spatial neighbors} \\ 0, \& \text{ else} \end{cases}$$

$$W = \sum_{i,j} w_{ij}.$$

Geary's *C*. Like Moran's *I*, Geary's *C* is also a metric to quantify the degree of spatial autocorrelation in spatial omics analysis. The difference is that Geary's C ranges from 0 to 2, while Moran's *I* ranges from −1 to 1. Following the same set of notations as Moran's *I*, the Geary's *C* is computed as follows. The relationship between Moran's *I* and Geary's *C* is explained in Supplementary Note 13.

$$\text{Geary's } C = \frac{N}{2W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i-x_j)^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

Divide and conquer strategy

We propose a divide and conquer approach to solve the large-scale spatial clustering problem as illustrated in Fig. 5. The original SpaceFlow is applied on the 31 slices independently. We adopted the procedure and parameter settings as suggested by the original author[81]. SpaceFlow can output both the context-aware embedding and spatial clustering labels. The spatial clustering labels is generated by Louvain clustering applied on the embedding. The embedding shown in Fig. 5a was generated by applying uniform manifold approximation and projection (UMAP) on the concatenated embeddings of all 31 slices to illustrate the existence of batch effect.

Harmony integration method[56] was applied to the concatenated embeddings (stored as an Anndata object in Python environment) of the 31 slices. We used the procedure suggested in SCANPY. In detail, the PCA was first applied on the concatenated embedding matrix to reduce the dimension to 30. Then Harmony integration function (SCANPY implementation) was applied on the Anndata object, with different slices indicated by 'slice_id'. Then a harmonized embedding was generated as stored in X_pca_harmony of the Anndata object. We built a neighborhood graph using the harmonized embedding, based on which UMAP visualization (as shown in the embedding of Fig. 5b) and Leiden clustering (as shown in the predicted label in Fig. 5b, right) can be performed.

After generating the harmonized embedding, clustering methods other than Leiden can be applied. We tested Louvain and mclust, together with Leiden, which are the most widely adpoted algorithms in existing spatial clustering methods. In the text and Fig. 5, SpaceFlow means the original SpaceFlow that can only be used for single-slice analysis. SpaceFlow-DC (or 'Proposed' or SpaceFlow-DC-mclust) means using the harmonized embedding followed by mclust to get spatial clustering labels. SpaceFlow-DC-Louvain means using the harmonized embedding followed by Louvain to get spatial clustering labels. SpaceFlow-DC-Leiden means using the harmonized embedding followed by Leiden to get spatial clustering labels. Embedding (Proposed) means using UMAP to visualize the harmonized embedding (Fig. 5b). Embedding (SpaceFlow) means using UMAP to visualize the nonharmonized embedding (Fig. 5a).

The per-slice NMI was recorded using the scikit-learn implementation of NMI to evaluate each slice's clustering accuracy, independently. The all-slice NMI was recorded to evaluate all cell's clustering accuracy across all 31 slices. Due to the SpaceFlow-predicted labels across slices are not aligned, the per-slice accuracy is high while the all-slice accuracy is low (Fig. 5f,g).

Computational resources

Intel Xeon E5-2683v3 central processing unit (CPU) (2.00 GHz, 35 MB L3 cache, 14 CPU cores in total), 128 GB memory, and NVIDIA TITAN Xp graphics processing unit (GPU) (12 GB of memory).

The rank-based overall score

In Fig. 4, we evaluated the accuracy, consistency, maker genes score and scalability of each method. As with previous benchmark studies[42], we defined the overall score for each component by integrating the metrics it contains (that is, accuracy contains NMI, HOM and COM; continuity contains CHAOS, ASW and PAS; marker genes contains Moran's *I* and Geary's C; scalability contains time and memory) to evaluate each methods among 10x Visium datasets or imaging-based datasets. As an example of the accuracy score based on imaging-based datasets, first, we calculated the average NMI, HOM and COM for each method among imaging-based datasets; then we sorted the NMI, HOM and COM values of methods in ascending order (from best to worst) to get rank_{NMI}, rank_{HOM} and rank_{COM}; finally, we calculated the average value of rank_{NMI}, rank_{HOM} and rank_{COM} to obtain the overall score value of accuracy, as follows:

$$\text{Overall score(accuracy)} = \frac{1}{3}\left(\text{rank}_{\text{NMI}} + \text{rank}_{\text{HOM}} + \text{rank}_{\text{COM}}\right)$$

The overall scores of others (that is consistency, maker genes score and scalability) are also used as described above, as follows:

$$\text{Overall score(consistency)} = \frac{1}{3}\left(\text{rank}_{\text{CHAOS}} + \text{rank}_{\text{ASW}} + \text{rank}_{\text{PAS}}\right)$$

$$\text{Overall score(maker genes score)} = \frac{1}{2}\left(\text{rank}_{\text{Moran's } I} + \text{rank}_{\text{Geary's } C}\right)$$

$$\text{Overall score(consistency)} = \frac{1}{2}\left(\text{rank}_{\text{Time}} + \text{rank}_{\text{Memory}}\right)$$

Correlation of methods performance

To assess the correlation between the relative performance of all methods applied to two data (for example, Data1 and Data2), we utilized Spearman's rank correlation coefficient implemented in the scipy package[82]. For the NMI metric (as reported in Fig. 3a), we formed two vectors to input into the Spearman's correlation computation. The first vector contained each method's NMI score on Data1, with each element representing the median NMI of ten independent runs. The second vector contained each method's NMI score on Data2, with each element representing the median NMI of ten independent runs. When assessing the relative performance between two spatial technologies

(as reported in Fig. 3b), a similar approach was adopted, with the only difference being that the performance vector was formed by taking the median scores across all data within the spatial technology.

Scientific computing and plotting

We used Python (version 3.9) environment to conduct this study. Important packages include scanpy[33] (version 1.9.1), numpy[83] (version 1.22.4), squidpy[84] (version 1.2.3), pandas[85] (version 1.5.1), scipy[82] (version 1.9.3), scikit-learn[86] (version 1.1.2), matplotlib[87] (version 3.6.0), seaborn[88] (version 0.12.1) and palettable[89] (version 3.3.0).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data1 to Data12 were downloaded from ref. 64. Data13 to Data21 are available from ref. 65. Data22 to Data24 were downloaded from ref. 66. Data25 to Data29 were downloaded from ref. 67. Data30 was downloaded from ref. 68. Data31 to Data33 are available from ref. 69. Data34 was downloaded from ref. 69. Data35 to Data41 were downloaded from ref. 70. Data42 to Data54 were downloaded from https://www.livercellatlas.org/. Data55 to Data56 are available at GSE111672. Data57 to Data87 were downloaded from ref. 71. Source data are provided with this paper.

Code availability

The code and scripts used for data preprocessing and visualization are available at https://github.com/zhaofangyuan98/SDMBench. Our benchmarking workflow is provided as a reproducible pipeline at https://github.com/zhaofangyuan98/SDMBench/tree/main/SDMBench. We also provide a tutorial at https://github.com/zhaofangyuan98/SDMBench/tree/main/Tutorial.

References

59. Wolf, F. A. et al. Louvain usage in Scanpy. *Scanpy* https://scanpy.readthedocs.io/en/stable/generated/scanpy.tl.louvain.html (2018).
60. Wolf, F. A. et al. Leiden usage in Scanpy. *Scanpy* https://scanpy.readthedocs.io/en/stable/generated/scanpy.tl.leiden.html (2018).
61. Hao, M., Hua, K. & Zhang, X. SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab471 (2021).
62. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).
63. Sun, S. et al. SPARK usage for spatially variable gene detection. *Xiang Zhou Lab* https://xzhoulab.github.io/SPARK/ (2020).
64. Maynard, K. R. et al. spatialLIBD for hosting dorsolateral prefrontal cortex 10x Visium dataset. *spatialLIBD* http://research.libd.org/spatialLIBD (2021).
65. Xu, Z. et al. STOmicsDB database page of mouse embryo Stereo-seq dataset. *China National GeneBank* https://db.cngb.org/stomics/mosta/ (2022).
66. Long, B. et al. Webpage of SpaceTx. *The SpaceTX Consortium* https://spacetx.github.io/ (2023).
67. Moffitt, J. R. et al. Data from: Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. *Dryad*. https://doi.org/10.5061/dryad.8t8s248 (2018).
68. Codeluppi, S. et al. Data and code availability. Expression data: loom file with osmFISH data. *Linnarsson Lab* http://linnarssonlab.org/osmFISH/availability/ (2018).
69. Wang, X. et al. Data from: Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Deisseroth Lab* http://clarityresourcecenter.org/ (2018).
70. Andersson, A. et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Zenodo* https://doi.org/10.5281/zenodo.4751624 (2021).
71. Allen, W. E. et al. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *CZ CELLxGENE* https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e (2022).
72. Wang, J. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* **12**, 1882 (2021).
73. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
74. Wang, B., Zhu, J. J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
75. Pedregosa, F. et al. Homogeneity score usage in scikit-learn. *scikit-learn* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html (2014).
76. Pedregosa, F. et al. Completeness score usage in scikit-learn. *scikit-learn* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html (2014).
77. Alexandrov, T. & Bartels, A. Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics* **29**, 2335–2342 (2013).
78. Guo, L. et al. Data filtering and its prioritization in pipelines for spatial segmentation of mass spectrometry imaging. *Anal. Chem.* **93**, 4788–4793 (2021).
79. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
80. Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X. & Fan, J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res.* **31**, 1843–1855 (2021).
81. Ren, H. et al. SpaceFlow. *GitHub* https://github.com/hongleir/SpaceFlow (2022).
82. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
83. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
84. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* https://doi.org/10.1038/s41592-021-01358-2 (2022).
85. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* Vol. 445 (eds van der Walt, S. & Millman, J.) 51–56 (2010).
86. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
87. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
88. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
89. Davis, M., Sick, J. & Eschbacher, A. palettable: color palettes for Python. *Astrophysics Source Code Library* ascl: 2202.2005 (2022).

Acknowledgements

This study was supported by National Nature Science Foundation of China (62303119, Z.Y.), Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (22CGA02, Z.Y.), Shanghai Science and Technology Development Funds (23YF1403000 Z.Y.), Tencent AI Lab Rhino-Bird Focused Research Program (RBFR2023008, Z.Y.), Innovation Fund of Institute of Computing and Technology, CAS (E161080 and E161030, Yi Zhao) and Beijing Natural Science Foundation Haidian Origination and Innovation Joint Fund (L222007, Yi Zhao). This work was also supported by Shanghai Municipal Science and Technology Major

Project (no. 2018SHZDZX01), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology, and 111 Project (no. B18015). The authors would like to acknowledge the Nanjing Institute of InforSuperBahn MLOps for providing the training and evaluation platform.

Author contributions

Yi Zhao and Z.Y. conceived and designed the study. Z.Y. and Yi Zhao designed the metrics, benchmark pipeline, and collected the methods and datasets. F.Z. and Z.Y. implemented the benchmarking pipeline. Z.Y. implemented the divide and conquer strategy. Z.Y. and F.Z. analyzed the results and generated the figures. Z.Y., F.Z. and Yi Zhao wrote the manuscript. Yu Zhao, J.Y. and Y.C. helped implement the large data scalability. X.Z. and J.Y. provided tissue anatomical knowledge. S.L. helped re-implement the methods.

Competing interests

The author declares no competing interests.

Additional information

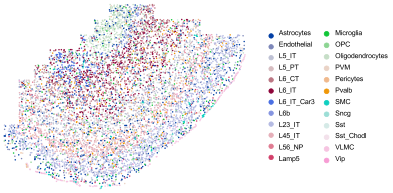
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02215-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02215-8.

Correspondence and requests for materials should be addressed to Zhiyuan Yuan or Yi Zhao.

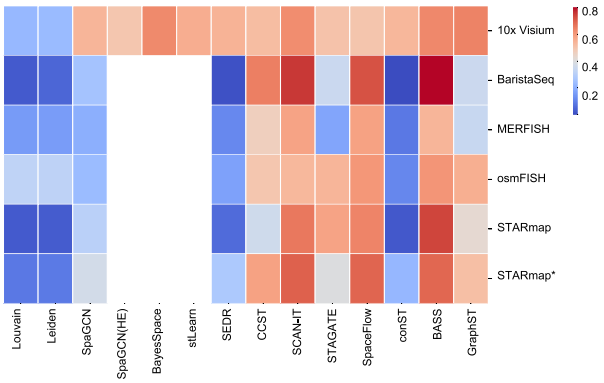
Peer review information *Nature Methods* thanks Karoline Holler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the Nature Methods team.

https://doi.org/10.1038/s41592-024-02215-8

Clustering Type	Goal	Feature	Representative work	Example
Spatial clustering	Partition tissues into distinct compartments so that each compartment exhibits spatially coherent gene expression pattern.	The feature used for clustering typically contains the spatial context information around the target cell.	SpaGCN (Hu et al., *Nature Methods* 2021) BayesSpace (Zhao et al., *Nature Biotechnology* 2021)	
Cell-type clustering	Identify cell populations so that cells with similar gene expressions belong to the same population.	The feature used for clustering typically contains the gene expression information of the target cell.	Louvain (Blondel et al., *Journal of Statistical Mechanics,* 2008) SC3 (Kiselev et al., *Nature Methods* 2017)	

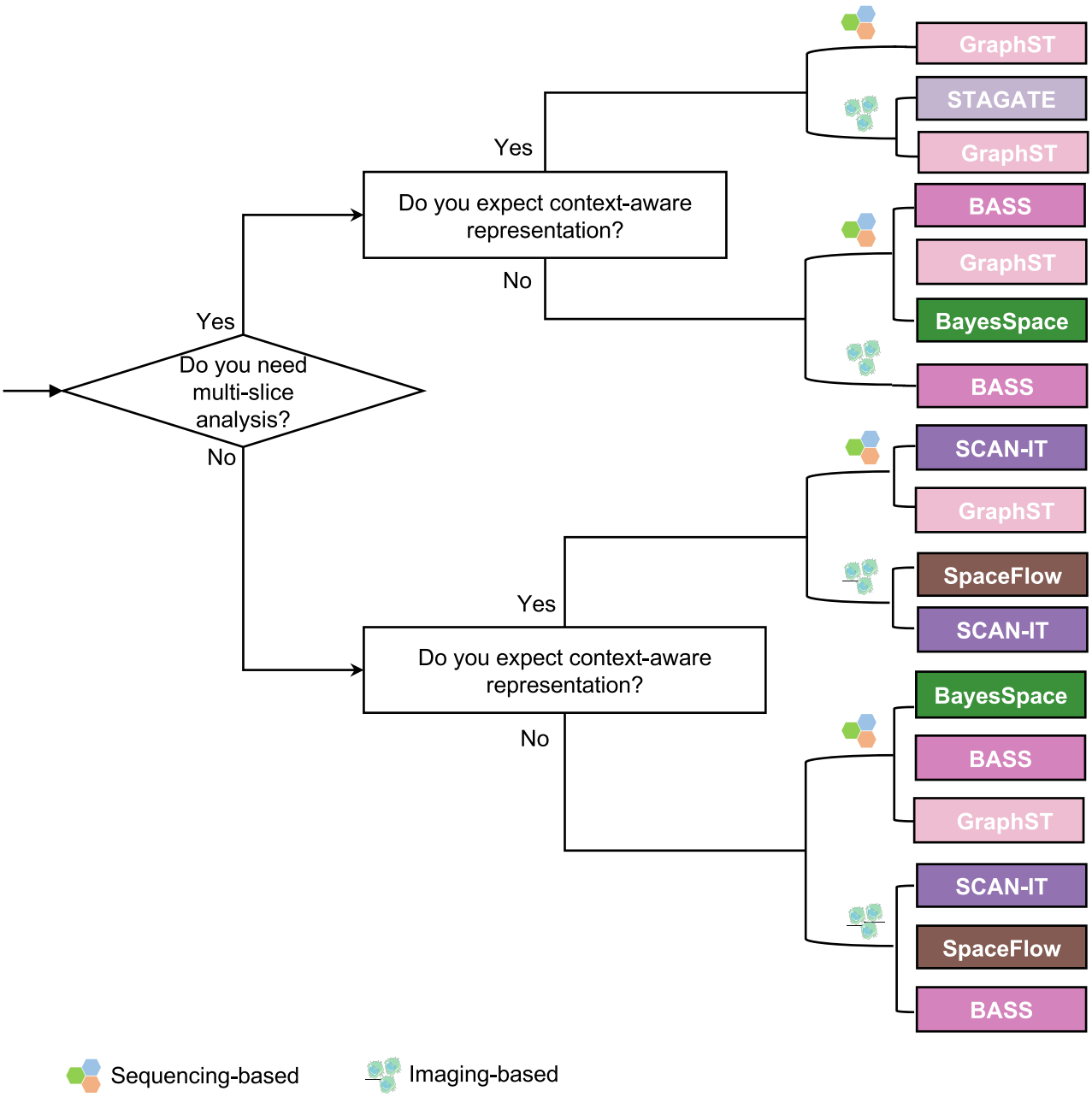
Extended Data Fig. 1 | The differences between spatial clustering and cell type clustering. Spatial clustering and cell type clustering are different tasks, we explained their differences in their goals, features, and representative work. We also used an example from mouse motor cortex data to explain their differences.

https://doi.org/10.1038/s41592-024-02215-8



Extended Data Fig. 2 | Methods performance on various biotechnologies. On the heatmap, the rows represent the biotechnologies, the columns represent the methods, and each value in the figure represents the NMI values.

https://doi.org/10.1038/s41592-024-02215-8



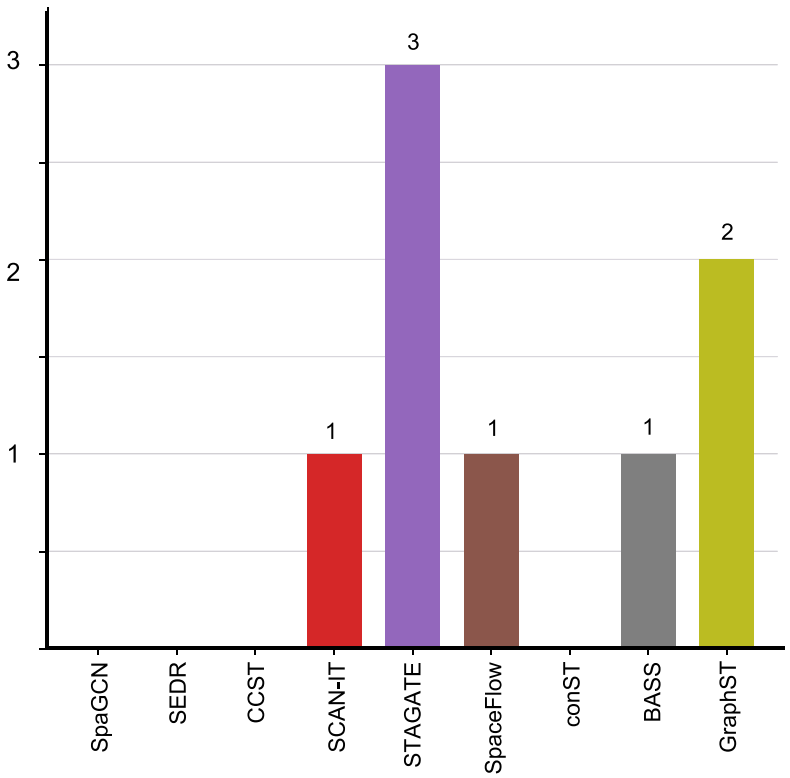
Extended Data Fig. 3 | User guidance. Recommend the suitable methods for users according to the data at hand. Note that the method choice was based on the accuracy scores. For more specific recommendations, users should look at Fig. 4 to refer to other aspects of performance.

https://doi.org/10.1038/s41592-024-02215-8

A

	SpaGCN	SEDR	CCST	SCAN-IT	STAGATE	SpaceFlow	conST	BASS	GraphST
Data 35	0.035	0.038	0.031	0.021	0.125	0.027	0.032	0.008	0.222
Data 36	0.060	0.088	0.094	0.105	0.462	0.091	0.068	0.034	0.263
Data 37	0.053	0.020	0.028	0.038	0.038	0.067	0.061	0.022	0.024
Data 38	0.046	0.094	0.047	0.054	0.182	0.106	0.069	0.041	0.122
Data 39	0.027	0.019	0.021	0.062	0.018	0.052	0.023	0.018	0.017
Data 40	0.081	0.106	0.087	0.107	0.078	0.097	0.089	0.075	0.087
Data 41	0.060	0.086	0.048	0.117	0.097	0.304	0.092	0.024	0.050
Data 42	0.066	0.054	0.036	0.045	0.093	0.032	0.038	0.047	0.314
Data 43	0.064	0.054	0.025	0.033	0.065	0.038	0.086	0.027	0.487
Data 44	0.035	0.055	0.033	0.034	0.070	0.076	0.042	0.115	0.429
Data 45	0.097	0.070	0.060	0.053	0.096	0.051	0.055	0.046	0.129
Data 46	0.071	0.089	0.055	0.040	0.095	0.042	0.058	0.042	0.186
Data 47	0.337	0.232	0.227	0.195	0.232	0.225	0.310	0.161	0.392
Data 48	0.167	0.139	0.121	0.099	0.132	0.152	0.220	0.064	0.131
Data 49	0.043	0.029	0.037	0.030	0.044	0.051	0.031	0.028	0.115
Data 50	0.411	0.331	0.277	0.538	0.616	0.435	0.281	0.221	0.733
Data 51	0.043	0.030	0.038	0.043	0.108	0.190	0.050	0.444	0.065
Data 52	0.383	0.498	0.374	0.303	0.660	0.127	0.494	0.617	0.396
Data 53	0.090	0.261	0.176	0.191	0.446	0.583	0.132	0.200	0.630
Data 54	0.468	0.361	0.293	0.392	0.617	0.451	0.279	0.245	0.408

B



Extended Data Fig. 4 | Performance on challenging datasets. A: This figure records all methods IoU across small and non-continuous data, where data35-data41 are breast cancer data and data42-data54 are liver data. B: This figure records the number of successful identifications (IoU >= 0.5) for each method.

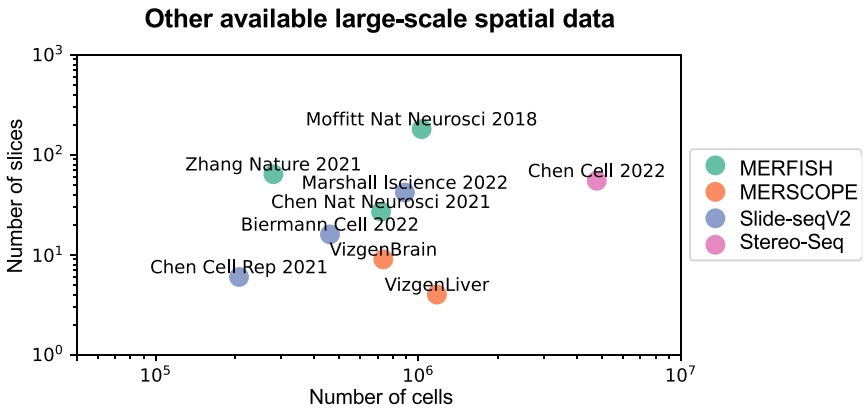
https://doi.org/10.1038/s41592-024-02215-8

A

Dataset information

Key	Value
Technology	MERFISH
# slices	31
# cells	378,918
# genes	374
Tissue type	frontal cortex & striatum
Resolution	single cell

B

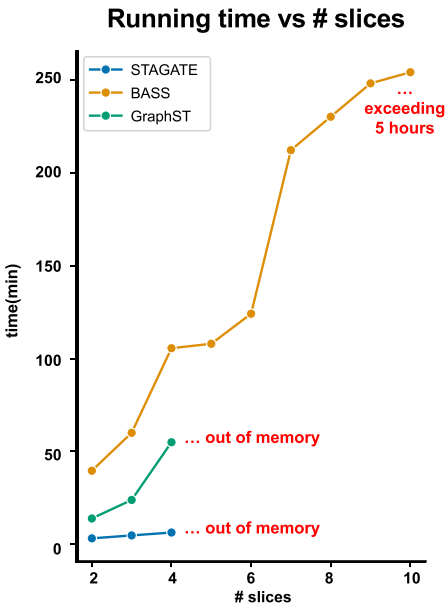


C

Issues of different methods when applied on (Allen et al. Cell 2022)

Method	Spatial information not considered	H&E image required	Spot-like input required	Time issue	Memory issue	Multi-slice analysis not supported
Louvain	■					
Leiden	■					
SpaGCN						■
SpaGCN (HE)		■				■
BayesSpace			■	■	■	
stLearn		■			■	■
SEDR					■	■
CCST					■	■
SCAN-IT					■	■
STAGATE					■	
SpaceFlow					■	■
conST					■	■
BASS				■		
GraphST					■	

D



Extended Data Fig. 5 | Limitations of current methods on large-scale datasets. A large-scale MERFISH dataset was used to illustrate that current methods cannot be applied on the dataset. A: The dataset information. B: Other large-scale datasets available in the field. Each point is a dataset, x stands for the number of cells, y stands for the number of slices. The publication information is annotated beside the points. Colors indicate different spatial technologies. C: Issues of each method when applied on the dataset in A. Time issue means the running time exceeds 5 hours, and memory issue means the program report"out of memory" error. Computational resources can be found in Methods. D: The running time of BASS and STAGATE, as the function of the number of slices of the dataset in (A).

nature portfolio

Corresponding author(s): Zhiyuan Yuan & Yi Zhao

Last updated by author(s): Nov 20, 2023

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	Raw data are collected using Python3.9.
Data analysis	scanpy==1.9.1 squidpy==1.2.3 numpy==1.22.4 pandas==1.5.1 scipy==1.9.3 matplotlib==3.6.0 seaborn==0.12.1 palettable==3.3.0 scikit-learn==1.1.2

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

Datat1 - Data12 were downloaded from http://research.libd.org/spatialLIBD. Data13 - Data21 were available from https://db.cngb.org/stomics/mosta/. Data22 – Data24 were downloaded from https://spacetx.github.io/. Data25 – Data29 were downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248. Data30 was downloaded from http://linnarssonlab.org/osmFISH/availability/. Data31 – Data33 were available from http://starmapresources.com/data. Data34 was downloaded from http://starmapresources.com/data. Data35 - Data41 were downloaded from https://zenodo.org/record/4751624#.ZFz1OXZBxjU. Data42 – Data54 were downloaded from https://www.livercellatlas.org/. Data55 – Data56 was available at [GSE111672]. Data57 – Data87 were downloaded from https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not relevant.
Population characteristics	Not relevant.
Recruitment	Not relevant.
Ethics oversight	Not relevant.

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	All the data used are from public domains, and the sample size is determined in the original publication. The sample information is reported in the method section and in Figure 1.
Data exclusions	None data were excluded from the raw.
Replication	each experiments were replicated for 10 runs.
Randomization	We used random seeds for methods randomization.
Blinding	The algorithm developer and the data analyzer are the same person, so totally blinding is impossible.

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

Methods

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