nature portfolio

Peer Review File

Weighted sliced inverse regression for scalable supervised dimensionality reduction of spatial transcriptomics data

Corresponding Author: Dr Shila Ghazanfar

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors present wSIR, an extension of sliced inverse regression tailored to spatial transcriptomics data. The method introduces spatial weighting between slices to capture local spatial correlations and enables projection of non-spatial gene expression data into spatially-informed low-dimensional embeddings. The authors demonstrate its performance across simulated data, SeqFISH mouse embryo data, and Visium breast cancer datasets.

Overall, this manuscript presents an interesting and potentially useful method, but the current version needs additional comparison and validation. I have several major comments:

Major comments:
1. wSIR is described as supporting gene interpretation, clustering, trajectory analysis, and spatial prediction/discovery (as in Fig. 1d). However, trajectory analysis is not demonstrated in the manuscript, and clustering or spatial prediction are not compared with established methods in these domains. I suggest the authors either provide comparisons or clarify the rationale for not doing so, and discuss the relative strengths and limitations of wSIR in these tasks.
2. wSIR appears capable of integrating data from multiple spatial samples (e.g., different layers of the same tissue). However, it is unclear how such integration improves the final embedding, and what assumptions or constraints are required (e.g., shared tissue origin, alignment of coordinate systems). Clarifying how cross-sample integration affects the learned representation would be valuable.
3. The method uses grid-based slicing, with fixed-size tiles and mean gene expression per slice used in the downstream decomposition. While this design is reasonable for relatively homogeneous tissues, its suitability for more heterogeneous tissues—such as cancer samples—need further investigation. In such cases, local spatial structure and expression heterogeneity might be obscured by within-slice averaging.
4. UMAP is a widely adopted nonlinear embedding method in ST and scRNA analyses. It would be helpful to include UMAP in comparative plots or at least discuss why it was omitted.
Minor comments:
1. The font sizes the figures are very small.
2. Fig. 2f, it would be better if presents the raw jaccard index rather that the ratio of PCA results.
3. Fig. 3c is a little bit hard to differentiate the results between wSIR and PCA.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
In this manuscript, Woollard et al. present a spatially-aware dimension reduction method (wSIR) for spatial transcriptomics data. The algorithm initiates by discretizing spatial coordinates into slices, which reduces dimensionality via slice-level gene expression aggregation while simultaneously modeling inter-slice relationships using a spatial proximity-based weight matrix. By integrating classical sliced inverse regression with spatial modeling, wSIR balances computational efficiency with spatial information preservation. The resulting gene loading matrix can be projected back to the original resolution, enabling effective analysis of large-scale ST datasets. Additionally, wSIR facilitates the projection of non-spatial (e.g., scRNA-seq) data into a spatially-informed low dimensional space. This linearly scalable architecture is particularly suited for high-throughput, subcellular-resolution spatial transcriptomics studies.
The authors evaluated wSIR through simulations and two biological case studies (human breast cancer and early mouse embryonic development). I have several concerns regarding benchmarking and evaluation rigor. My main comments are as follows:

Major comments:
1. (Materials and Methods) The performance of wSIR critically depends on the choice of the number of slices (H). While the current equal-interval slicing approach offers computational simplicity, I have concerns about its biological appropriateness for heterogeneous tissues. For instance, fixed-size slicing may artificially aggregate distinct cellular niches (e.g., at the tumor-immune border), potentially diluting biologically meaningful variation. Conversely, overly fine slicing could undermine computational efficiency and resemble PCA behavior. The authors should explore or discuss adaptive slicing strategies, e.g., pre-clustering cells/spots based on both spatial proximity and transcriptional similarity. Furthermore, the authors should provide practical guidance for selecting hyperparameters (e.g., number of tiles H and spatial weighting coefficient α) across different SRT technologies and resolutions, which could enhance the method's practical utility.

2. (Simulation desigan) The current simulation framework does not fully reflect the spatial continuity inherent in real ST data. First, the current linear framework fails to account for the characteristic of spatial transcriptomics data where neighboring spots typically exhibit similar gene expression patterns. This oversimplification may artificially inflate wSIR's apparent performance. Second, the simulation should incorporate more biologically realistic count distributions (e.g., negative binomial or Poisson) and spatial autocorrelation. Third, the evaluation would benefit from including more diverse simulation scenarios such as spot sparsity, and noise levels.

3. (Running time and memory usage comparison). While wSIR is claimed to be computationally efficient, the current evaluation lacks systematic benchmarking of runtime and memory usage across datasets.

4. (Method comparison). As a dimensionality reduction method specifically developed for spatial transcriptomics, wSIR should be systematically benchmarked against spatially-aware methods (e.g., SpatialPCA1, DR-SC2, PRECAST3, and STAMP4) rather than predominantly spatially-unaware methods (e.g., PCA and LDA). This is especially critical for the real-data analyses, where GraphPCA is the only spatially-aware baseline considered.

5. (Evaluation metrics). While BCDC measures spatial concordance, it does not directly evaluate the discriminative power of the embeddings for cell type or domain prediction. Incorporating additional metrics like the McFadden-adjusted pseudo-R^2 (as used in SpatialPCA1) would provide a more rigorous assessment of the embeddings' predictive capacity for biological labels. Furthermore, the manuscript should systematically assess the biological relevance of the low-dimensional embeddings by benchmarking performance on critical analytical tasks such as spatial domain detection/cell type clustering using metrics like the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

6. (Spatial prediction evaluation). The use of wSIR embeddings in Tangram demonstrates acceleration in spatial mapping, but spatial reconstruction accuracy should also be quantitatively assessed. Incorporating measurements such as Jensen-Shannon divergence and Spearman's rank correlation (as implemented in CellContrast5) would provide a more comprehensive and interpretable evaluation.

Minor comments:
1. (Introduction, Line58) MUSTARD is a trajectory-guided dimensionality reduction method for scRNA-seq data, not specifically designed for spatial transcriptomics data.
2. (Results) For the mouse embryo analysis, the authors should report post-QC cell numbers for all three replicates and include spatial plots of ground truth and annotation results by each method.
3. (Figure 3) Please increase font size in Figure 3. The current text is too small to read clearly.
4. (Figure 3h) The legend for dot sizes is partially truncated.
5. (Figure 3i) The x-axis label 'WSIR' should be 'wSIR'; The x-axis title 'DR method' should be 'Input type of Tangram'.
6. (Evaluation metrics) The methodological details of mean-squared error (MSE) require explicit documentation. Besides, the authors should provide spatial visualizations of reconstructed locations.
7. (Format) Adjust paragraph format following equations. When an equation concludes (e.g., Lines 360 and 371), the subsequent paragraph should begin with an indentation.

References
1. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. Nat Commun 13, 7203 (2022).
2. Liu, W. et al. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. Nucleic Acids Res 50, e72 (2022).
3. Liu, W. et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. Nat Commun 14, 296 (2023).
4. Zhong, C., Ang, K.S. & Chen, J. Interpretable spatially aware dimension reduction of spatial transcriptomics with STAMP. Nat Methods 21, 2072-2083 (2024).
5. Li, S. et al. CellContrast: Reconstructing spatial relationships in single-cell RNA sequencing data via deep contrastive learning. Patterns (N Y) 5, 101022 (2024).

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors develop a modification of SIR to better retain the predictive power for the spatial coordinates. More precisely, the slicing step treats the spatial coordinates as multiple responses, and notably, correlations among slices are incorporated into the kernel matrix M. Both of simulation results and case studies indicate that the method yields promising performance across various scenarios, and outperforms existing alternatives. Overall, the manuscript is well written, and the methodology is clearly described and easy to follow.

While the proposed method appears well-motivated and demonstrates promising performance in the current manuscript, a few concerns remain regarding its theoretical foundation and potential limitations. Some issues are listed below:

* Theoretical foundation: The manuscript would benefit from a clearer discussion or formal analysis of the method's theoretical properties, such as the consistency of the estimators. Even partial theoretical results or heuristic arguments could help clarify the method's behavior.

* Validity of surrogate dependence structure: The use of distance correlation to represent the underlying spatial structure is a reasonable choice, but it would be helpful if more justification for this surrogate could be provided.

* Risk of false discoveries: Although the method seems effective at detecting meaningful spatial patterns in the presented examples, it would be important to evaluate its behavior under null scenarios. For example, when the features are unrelated to the spatial coordinates, would the method still produce spurious signals?

* Clarification of limitations: It is helpful to discuss the limitations of the proposed method, such as scenarios where it might seem applicable but is actually unsuitable. This could guide practitioners in avoiding potential misuse.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors' efforts. All my concerns have now been addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I appreciate the authors' efforts in addressing my previous concerns. However, there are still several specific points require further clarification.

1. While geometric invariance is a valuable mathematical property, evaluating the algorithm's robustness against varying spot sparsity and noise/dropout levels remains critical. Because wSIR relies heavily on summarizing data within spatial slices, its intra-slice covariance estimations are theoretically highly vulnerable to severe sparsity. Therefore, I suggest the authors perform the sparsity/dropout level simulations to assess the method's robustness.

2. For the DLPFC dataset, the median Adjusted Rand Index (ARI) for wSIR falls below 0.3. More concerning is that the performance of the competing algorithms reported in this study is significantly lower than benchmark results reported by Hu et al.1. The authors must carefully check code and hyperparameter settings for these competing methods to ensure a fair and accurate evaluation. Furthermore, the authors should provide exact P-values (e.g., via paired Wilcoxon signed-rank tests) for the comparisons in Supplementary Figure 6a and 6b.

3. The authors emphasize scalability as a primary superiority of wSIR. While Supplementary Figure 1 shows favorable runtime and memory usage, I am not fully convinced of its practical superiority based on the current presentation. Only a small subset of baseline methods was included in the comparison for Figures 2 and 3. In Figure 4, no other spatially-aware baseline methods were tested.

4. The reported mean Spearman correlation for wSIR in the Tangram spatial mapping task is only 0.129. Such a low absolute correlation value raises questions about the actual biological utility of the spatially-informed embeddings for reconstruction.

Minor points:
1. The font sizes in Figure 2 and 3 are still too small to read clearly.

Hu, Y. et al. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. Genome Biol 25, 212 (2024).

(Remarks on code availability)
I did not checked the code, but the performance of the competing algorithms reported in this study seems too low.

Reviewer #4

(Remarks to the Author)
I thank the authors for their detailed response letter and the revision made to the manuscript. While my previous concerns have been addressed to some extent, several points would still benefit from further clarification and more explicit discussion in the revised manuscript. In this report, I therefore provide more specific follow-up questions and suggestions to help clarify the scope, robustness, and practical use of wSIR. My main comments are listed below.

1 (Theoretical foundation) I agree with the authors' statement that "wSIR is a useful tool for computational biology". Thus, it is important to clarify the scenarios and underlying assumptions under which wSIR is expected to be useful, and when it may be less reliable. While a full theoretical analysis may be considered out of scope, it would be helpful to provide a more theoretical discussion of when and why wSIR is expected to work well.

2 (Validity of surrogate dependence structure) I appreciate the authors' clarification regarding the motivation for using bias-corrected distance correlation, as well as the inclusion of additional quantitative metrics in the revised manuscript. I would like to clarify that my concern is not about these evaluation metrics, but rather about the working dependence structure used in wSIR,
$w_{ij} = (1-d_{ij}/d_{max})^{(1/alpha)}$.
Notably, this form enforces $w_{ij} = 0$ for the longest-distance pair(s) with $d_{ij}=d_{max}$, regardless of alpha, which may be restrictive when long-range spatial dependence is present. To strengthen the manuscript, it would be helpful if the authors could either (i) provide a sensitivity analysis under misspecified dependence structures (e.g., exponential or Matern-type correlations with non-negligible long-range dependence), or (ii) discuss a more flexible framework that allows alternative correlation structures, together with practical guidelines for choosing an appropriate structure in applications.

3 (Risk of false discoveries) I appreciate the authors' permutation experiment in the response letter. The reduced leading eigenvalues under this null setting are informative. It would be helpful if the authors could describe this permutation-based procedure more explicitly and provide guidance on how practitioners can use it to assess whether the extracted components reflect real spatial signal.

4 (Clarification of limitations) I appreciate the authors' added discussion on the potential limitation that wSIR may be less effective at identifying non-spatial but still biologically relevant variation. In addition to the robustness issues discussed above regarding the working dependence structure, it would be helpful to further clarify the scope of applicability by discussing a few concrete scenarios where the method may be less appropriate. For example, it would be helpful to

comment on whether wSIR remains reliable when some covariates are categorical. If not, it may be useful to briefly discuss whether extensions such as partial dimension reduction (see, e.g., Li, Cook, and Chiaromonte, 2003) could be adapted to this setting.

5 (Minor) I noticed that Figure 2 has been updated substantially, but its caption still appears to correspond to the previous version. I would recommend that the authors to carefully proofread the revised manuscript to ensure that all elements (figures, captions, etc.) are consistent with the updated content.

References:
Li, B., Cook, R. D., & Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. The Annals of Statistics, 31(5), 1636-1668.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
My concerns have been addressed. I have no further questions.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I appreciate the authors' efforts in revising the manuscript. My previous concerns have been largely addressed.

At this stage, I have only one remaining suggestion. The comparison of different dependence structure models discussed in the response to Comment 2 (particularly the figure and the corresponding explanation) is quite informative. I recommend incorporating these results into the manuscript, as they provide more direct and intuitive evidence supporting the advantage of the proposed method.

Other than this, I have no further comments. The manuscript is now in good shape.

(Remarks on code availability)

We sincerely thank the anonymous reviewers and editor for their thoughtful and constructive feedback. As a result, we have revised the manuscript and included a point-by-point response to each reviewer's comments.

In this document, black text represents original reviewer comments, blue text represents author responses, and *magenta italics text* represents additions to the revised manuscript.

Reviewer: 1
Comments to the author:

The authors present wSIR, an extension of sliced inverse regression tailored to spatial transcriptomics data. The method introduces spatial weighting between slices to capture local spatial correlations and enables projection of non-spatial gene expression data into spatially-informed low-dimensional embeddings. The authors demonstrate its performance across simulated data, SeqFISH mouse embryo data, and Visium breast cancer datasets.

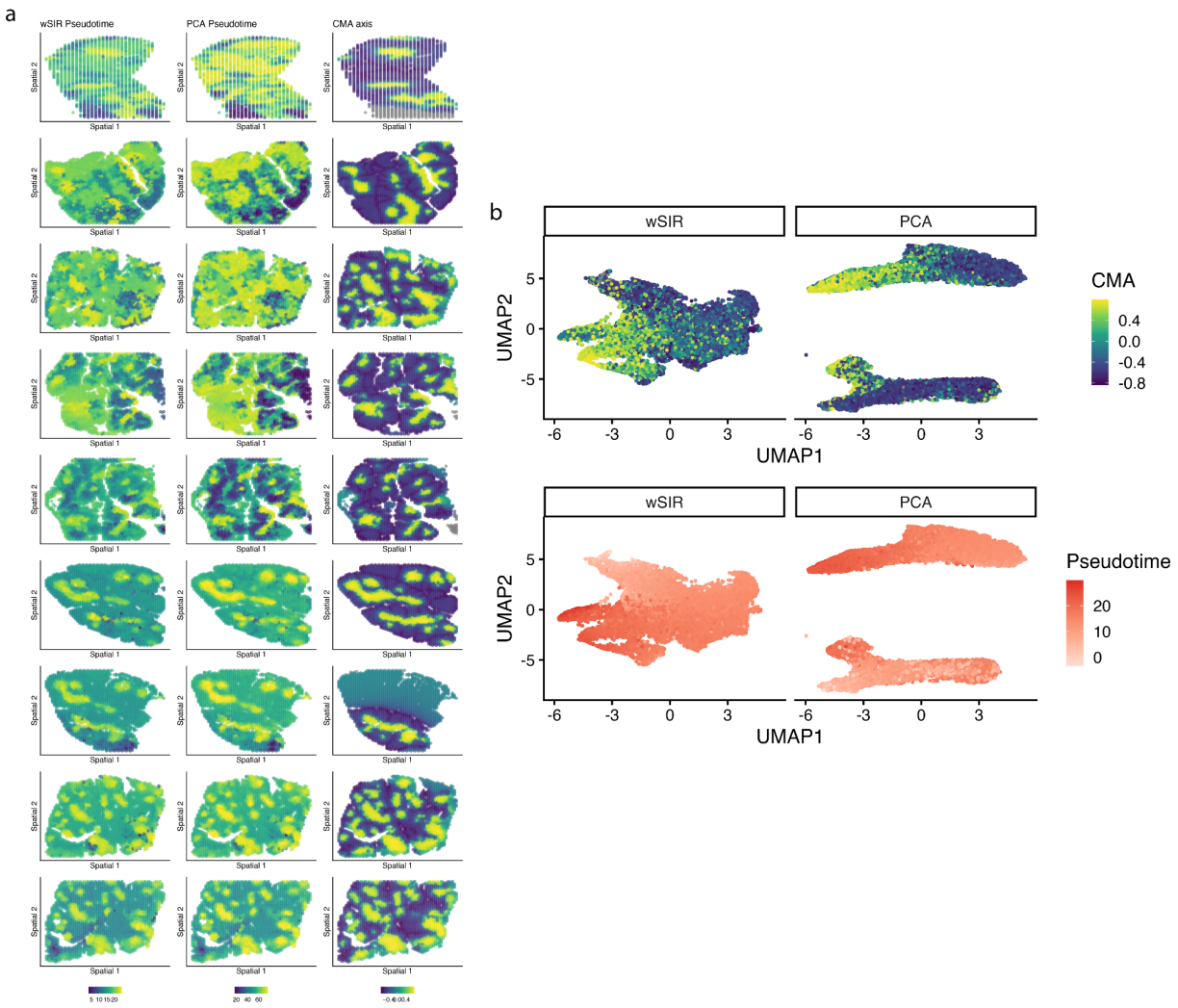
Overall, this manuscript presents an interesting and potentially useful method, but the current version needs additional comparison and validation. I have several major comments:

Thank you for the positive comment on our work.

Major comments:

1. wSIR is described as supporting gene interpretation, clustering, trajectory analysis, and spatial prediction/discovery (as in Fig. 1d). However, trajectory analysis is not demonstrated in the manuscript, and clustering or spatial prediction are not compared with established methods in these domains. I suggest the authors either provide comparisons or clarify the rationale for not doing so, and discuss the relative strengths and limitations of wSIR in these tasks.

Thank you for this feedback. Indeed, while we described that trajectory analysis, clustering, and spatial prediction are key downstream analysis tasks made possible with wSIR, we did not explicitly demonstrate these. To demonstrate trajectory analysis in pseudospace, we have now performed an additional analysis using a human thymus dataset that contains an annotated Cortico-Medullary Axis (CMA)[1]. We first performed wSIR and PCA on all 9 samples together, then removed any batch effect from the embeddings by applying the `RunHarmony()` function from the Harmony R package. Next, we estimated the pseudotime on the wSIR and PCA embeddings of each sample by using the `slingshot()` function from the slingshot R package. We then compared the estimated pseudotime to the true CMA. Supplementary Figure 5 (copied below) displays the wSIR and PCA trajectories and the true CMA for each sample on the spatial coordinates. The estimated trajectories on the wSIR and PCA embeddings are both relevant to the CMA, therefore demonstrating that wSIR is capable of performing this task. Furthermore, UMAP visualisations of the embeddings (Supplementary Figure 5b) demonstrates that the estimated pseudotime from wSIR is comparable to that of PCA in terms of CMA. This discussion is now added to Section 2.6 along with the description of the human thymus data in Section 5.4 of the revised manuscript.



Supplementary Figure 5. a. Estimated pseudotimes from wSIR and PCA embeddings compared to true CMA axis coloured on spatial plots. Results shown for 9 samples from human thymus data. b. UMAP plots on wSIR and PCA embeddings coloured by estimated pseudotime and CMA values on the combined 9 samples.

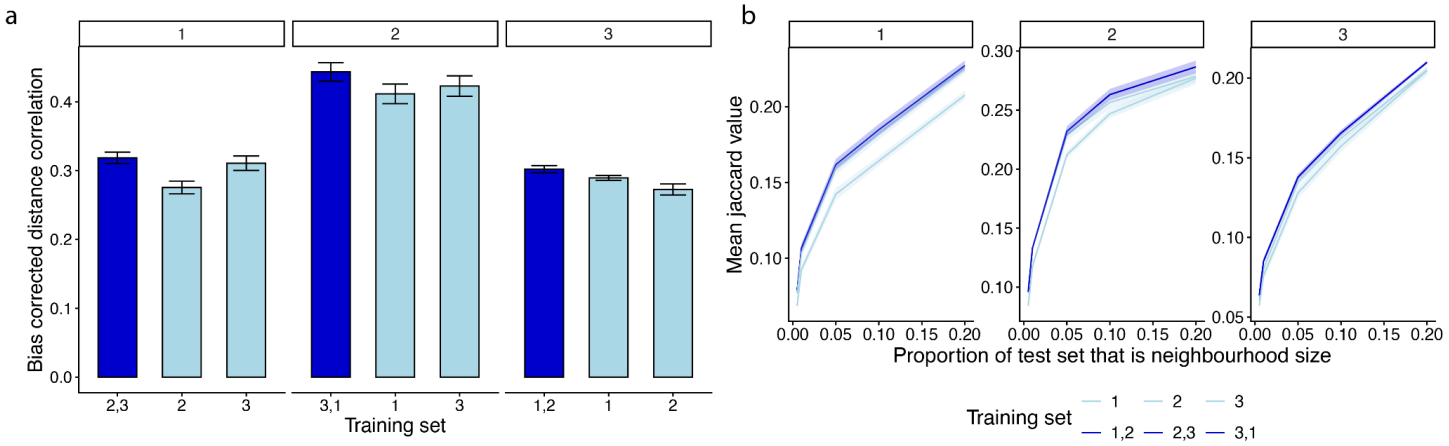
We have added the following text to the manuscript:

> *We next wondered if wSIR may be useful in uncovering axes of biologically relevant variation in space. To this end, we used a recently published human thymus spatial omics dataset (Yayon et al., 2024) where a continuous cortico-medullary axis (CMA) was defined. The CMA represents a continuous, rather than discrete, shift from tissue areas most associated with the cortex and medulla of the thymus. We used the nine spatial omics-resolved samples to perform wSIR dimension reduction alongside PCA, followed by trajectory 'pseudospace' inference using Slingshot (Street et al., 2018) (Methods). We observed that both approaches were able to recapitulate the CMA well, with Spearman correlations of 0.39 and 0.22 for wSIR and PCA respectively (Supplementary Figure 5a-b), and fairly close concordance between wSIR and PCA overall (Spearman correlation 0.57). Overall, this analysis demonstrates wSIR's ability to prioritise spatially relevant and continuous gene expression variation.*

Regarding comparisons of clustering and spatial prediction to other methods, we have now incorporated comparisons with other methods such as SpatialPCA, DR.SC, PRECAST, and STAMP, with full details provided below in responses to Reviewer 3 Comments 2, 5, and 6.

2. wSIR appears capable of integrating data from multiple spatial samples (e.g., different layers of the same tissue). However, it is unclear how such integration improves the final embedding, and what assumptions or constraints are required (e.g., shared tissue origin, alignment of coordinate systems). Clarifying how cross-sample integration affects the learned representation would be valuable.

Thank you for raising this point. We do empirically find that incorporating multiple samples improves the embedding. The new Supplementary Figure 4 (copied below) demonstrates that compared to when trained on only one sample, wSIR trained in two samples of the mouse embryo spatial transcriptomics data leads to higher bias-corrected distance correlations (BCDC) between the embedding and the spatial coordinates, as well as higher mean Jaccard values that measure neighbourhood similarities. We believe that this is because if the multiple samples have the same tissue origin, then this is effectively just adding more data points to the method. The only consideration with this is to ensure that wSIR does not see two points as being spatially similar if they have similar coordinates but belong to different samples. We achieve this by setting the weight matrix to 0 if that entry corresponds to tiles from different samples.



Supplementary Figure 4. a. BCDC performance metric of wSIR with 50 dimensions when trained on single or multiple samples (x-axis labels) of the mouse embryo dataset. b. Mean Jaccard index value for neighbourhood similarity between wSIR embeddings and spatial coordinates. The label on each panel denotes the sample that is used as the test set, while the line colour indicates the sample(s) used as the training set, darker blue indicates training with two samples.

We added the following discussion in Section 4.2 Applying wSIR to multiple samples:

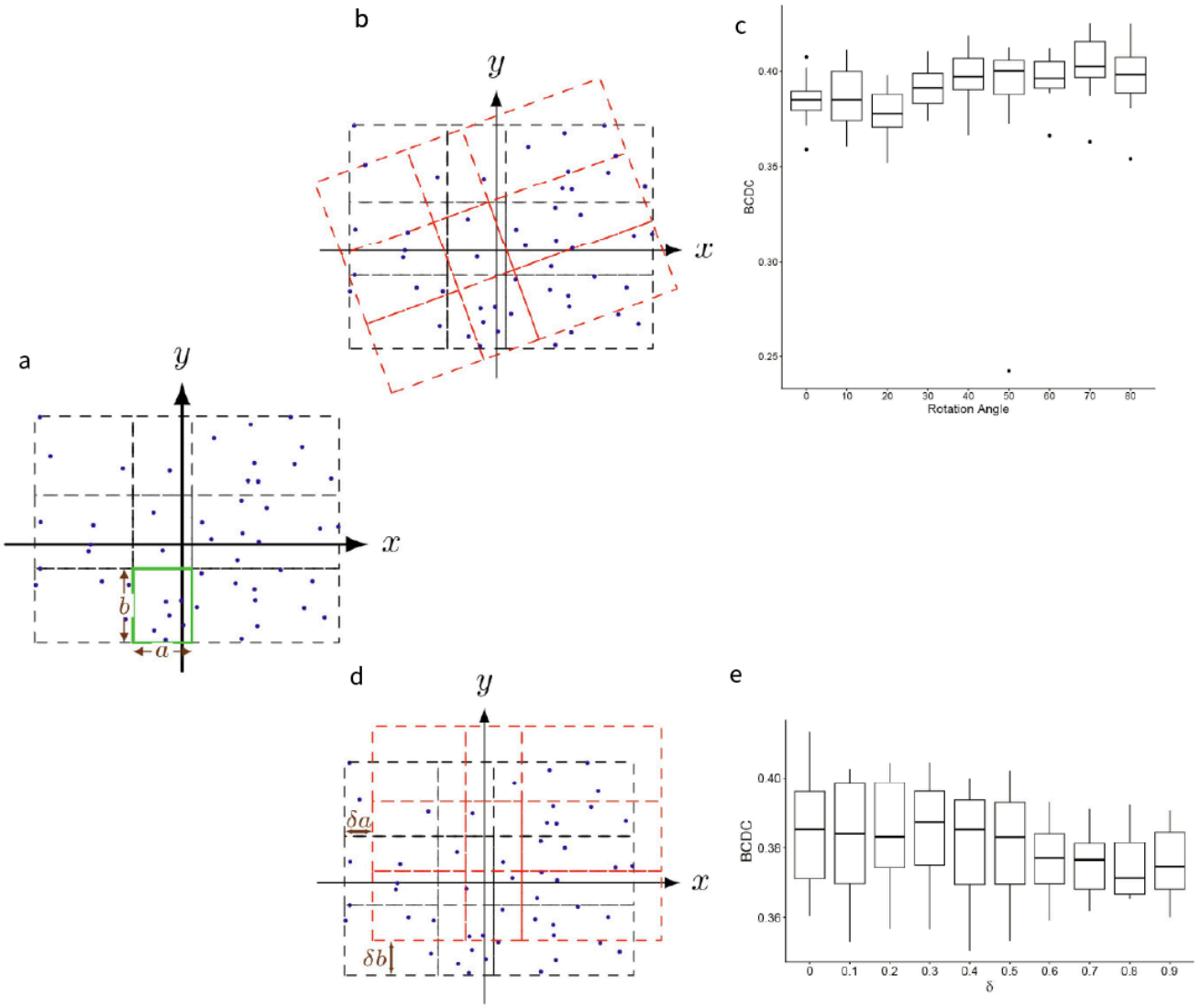
> *Because of the structure of the weight matrix W when wSIR is applied to multiple samples, there are few constraints on the samples included. Performing wSIR together would only add value if there is some spatially-relevant information that is common to all samples, so having shared tissue origin is required. The procedure also requires that the gene sets in all the samples are the same. Further, because wSIR does not include any batch correction stage, it is assumed that the input to wSIR will have already undergone batch correction. Note that we also do not need shared or aligned coordinate systems between the samples, because wSIR will automatically set the element in the weight matrix corresponding to two tiles from different samples to zero.*

3. The method uses grid-based slicing, with fixed-size tiles and mean gene expression per slice used in the downstream decomposition. While this design is reasonable for relatively homogeneous tissues, its suitability for more heterogeneous tissues—such as cancer samples—need further investigation. In such cases, local spatial structure and expression heterogeneity might be obscured by within-slice averaging.

This is a great point. Our intuition behind the choice of tile size and shape for wSIR is that it may be more sensitive to the tile *size* rather than the tile *shape* itself, which is also aligned with the existing literature on the traditional sliced inverse regression[2,3]. This is because, once we assign cells/spots to tiles, we perform aggregation followed by further eigendecomposition, so whether we assign very small or very large tiles may have the potential impact of obscuring any rare cells that are scattered across the tissue space. Indeed, in our investigations, we set the tile size as a key tuning parameter to be selected, while the tile shape is not investigated further.

To investigate the potential impact of tile shape, we performed an additional experiment on the xenium breast cancer dataset where we adjusted the spatial coordinates slightly in ways that should not modify the wSIR embedding. This aims to show that wSIR is not only suitable for the breast cancer dataset itself, as shown by its strong performance on this dataset compared to other methods, but also robust to heterogeneous cancer datasets in general.

The adjustments we make are rotating and shifting the grid structure that creates the tiles. For the rotation, we rotate the original grids that defines the spatial tiles by a certain angle. For the shift, we take the grid and shift it to the right and up by an amount defined by the size of the smallest tile. Supplementary Figure 2 (copied below) shows the boxplots for the BCDC corresponding to 10 subsamples of the Xenium Breast Cancer dataset, demonstrating that performance of wSIR is not sensitive to the considered levels of rotation and shifting respectively.



Supplementary Figure 2. Sensitivity analysis of wSIR to slicing mechanisms. (a), wSIR's default slicing mechanism is constructed on a rectangular grid of 2-D coordinate systems, (b), a slicing mechanism is formed by rotating the grid in panel a by a certain degree, (c), boxplots displaying bias-corrected distance correlation (BCDC) between wSIR's loadings and original spatial coordinates for 10 subsamples of the Xenium Breast Cancer dataset as a function of the rotation angle between the original grid and the grid in panel b. (d), a slicing mechanism is formed by shifting the grid in panel a to the right by δa and upward by δb,

where a and b are width and the height of the rectangle with smallest area in the grid of panel a, (e) same as panel c but for different values of δ for the shift in panel d.

4. UMAP is a widely adopted nonlinear embedding method in ST and scRNA analyses. It would be helpful to include UMAP in comparative plots or at least discuss why it was omitted.

We agree that UMAP is very popular and useful to visualise high-dimensional single-cell and ST data. However, there has been considerable discussion and debate on the potential pitfalls of using UMAP embeddings in a quantitative manner[4,5], as such, we have chosen to not include UMAP as a comparison for the embedding and any quantitative analyses that follow. Despite this, in line with the current state-of-the-art, we make qualitative use of UMAP to visualise the embeddings resulting from wSIR and PCA (e.g. in Figures 3 and 4), since wSIR and PCA routinely return embeddings in more than two dimensions with real data, and need further non-linear embedding to be effectively visualised in 2 dimensions.

Minor comments:

1. The font sizes of the figures are very small.

We have fixed the font sizes for readability.

2. Fig. 2f, it would be better if presenting the raw jaccard index rather that the ratio of PCA results.

Thanks for this, we have revised the Figure 2f to display the raw Jaccard index.

3. Fig. 3c is a little bit hard to differentiate the results between wSIR and PCA.

Apologies for this. We have clarified that the difference between the results of wSIR and PCA are that top genes associated with wSIR have localised expression in space.

Reviewer: 2

Comments to the author:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you for your work in co-reviewing our manuscript.

Reviewer: 3
Comments to the author:

In this manuscript, Woollard et al. present a spatially-aware dimension reduction method (wSIR) for spatial transcriptomics data. The algorithm initiates by discretizing spatial coordinates into slices, which reduces dimensionality via slice-level gene expression aggregation while simultaneously modeling inter-slice relationships using a spatial proximity-based weight matrix. By integrating classical sliced inverse regression with spatial modeling, wSIR balances computational efficiency with spatial information preservation. The resulting gene loading matrix can be projected back to the original resolution, enabling effective analysis of large-scale ST datasets. Additionally, wSIR facilitates the projection of non-spatial (e.g., scRNA-seq) data into a spatially-informed low dimensional space. This linearly scalable architecture is particularly suited for high-throughput, subcellular-resolution spatial transcriptomics studies.

The authors evaluated wSIR through simulations and two biological case studies (human breast cancer and early mouse embryonic development). I have several concerns regarding benchmarking and evaluation rigor. My main comments are as follows:

Thank you for your positive comments on our work.

Major comments:
1. (Materials and Methods) The performance of wSIR critically depends on the choice of the number of slices (H). While the current equal-interval slicing approach offers computational simplicity, I have concerns about its biological appropriateness for heterogeneous tissues. For instance, fixed-size slicing may artificially aggregate distinct cellular niches (e.g., at the tumor-immune border), potentially diluting biologically meaningful variation. Conversely, overly fine slicing could undermine computational efficiency and resemble PCA behavior. The authors should explore or discuss adaptive slicing strategies, e.g., pre-clustering cells/spots based on both spatial proximity and transcriptional similarity. Furthermore, the authors should provide practical guidance for selecting hyperparameters (e.g., number of tiles H and spatial weighting coefficient α) across different SRT technologies and resolutions, which could enhance the method's practical utility.

Thank you for this great point, which was also commented on by Reviewer 1. We agree that the number of slices (H) may play a large role in the wSIR embedding, as it is balancing the resolution to which observations are aggregated prior to the eigendecomposition step. The idea of adaptive slicing is a very interesting one that we did not investigate in the manuscript, though we note that existing implementations of SIR with continuous responses for non-spatial data usually use fixed-size slicing (see e.g. R packages dr[6], LassoSIR[7], Python package sliced[8]). In our revision, we have demonstrated that wSIR embeddings are robust to rotations and translocations of the tiles (please see response to Reviewer 1, Supplementary Figure 2), yet we have not performed any further investigation on adaptive slicing. We have added the following text to the discussion section of our revised manuscript:

> *We have shown that wSIR is quite robust to changes in the slicing mechanism, such as location shifts, but we have not fully examined the impact of other types of slicing mechanisms that, for example, take into account density of cells in a tissue. Given this, there is a future opportunity to investigate whether an 'adaptive' slicing mechanism may be beneficial in balancing between large enough sample sizes for analytical stability and small enough sample sizes to retain subtle spatial patterns.*

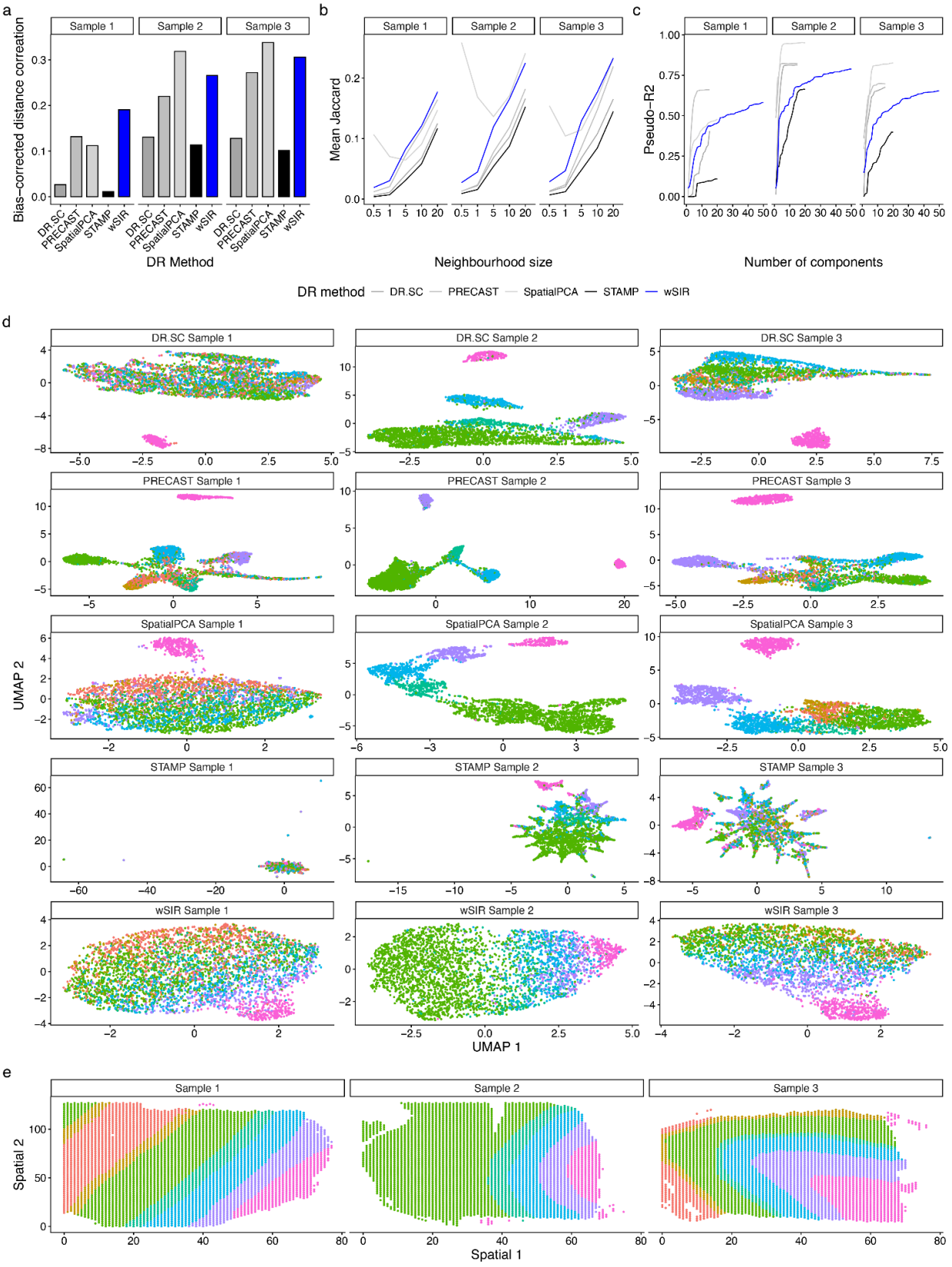
Regarding practical guidance for selecting hyperparameters, we include a hyperparameter tuning step in our software package that uses cross-validation to identify an optimal combination of H and alpha based on bias-corrected distance correlation (BCDC). Even so, we actually observe that wSIR is quite robust to specification of the hyperparameters. In Figures 2b and 2c, we performed wSIR with a range of parameter combinations, and found that wSIR performs quite well for any reasonable values of these parameters (close to default values).

2. (Simulation design) The current simulation framework does not fully reflect the spatial continuity inherent in real ST data. First, the current linear framework fails to account for the characteristic of spatial transcriptomics data where neighboring spots typically exhibit similar gene expression patterns. This oversimplification may artificially inflate wSIR's apparent performance. Second, the simulation should incorporate more biologically realistic count distributions (e.g., negative binomial or Poisson) and spatial autocorrelation. Third, the evaluation would benefit from including more diverse simulation scenarios such as spot sparsity, and noise levels.

Thank you for this comment. As far as we understand, the simulation framework that is being referred to here is the 'doughnut' that we describe in Figure 2a. In this case, we completely agree that it is an extremely simple simulated data that is created to illustrate that, while wSIR is a linear method, it is capable of identifying the linear combinations of data that are potentially *non-linearly* related to the outcome, that is, the 2-dimensional spatial coordinates. We believe that our benchmarking results performed on real data examples (Figures 2d-g, Supplementary Figure 6a-d) are strong evidence for the benefit of wSIR in prioritising spatially-relevant variation, rather than necessarily generating further simulation that may be unrealistic.

As a more realistic simulation dataset, we also compare wSIR against other spatial dimension reduction methods using the simulated data from the PRECAST vignette, available at https://feiyoung.github.io/PRECAST/articles/PRECAST.Simu.html. For this dataset, we first perform spatial dimension reduction methods on each of the three samples. We then compute the bias-corrected distance correlation between the embeddings and the true spatial coordinates, and the neighbourhood similarity between the embeddings and the true spatial coordinates (Supplementary Figures 7a,b). Using the true spatial domains, we also compute the pseudo-R^2 metric on the embeddings over a range of embedding sizes (supplementary figure 7c). To qualitatively assess the performance of the methods, we perform UMAP on the embeddings and colour by true spatial domains. These are shown in Supplementary Figure 7d, with the true spatial coordinates coloured by spatial domain included in Supplementary Figure 7e (copied below). These results demonstrate that wSIR is comparable with other considered methods, even in the tasks that wSIR is not designed for, i.e. when performing wSIR, we do not take clustering into account.

We have added this simulated dataset to Section 5.1 and the discussion regarding the results in Supplementary Figure 7 in Section 2.6.



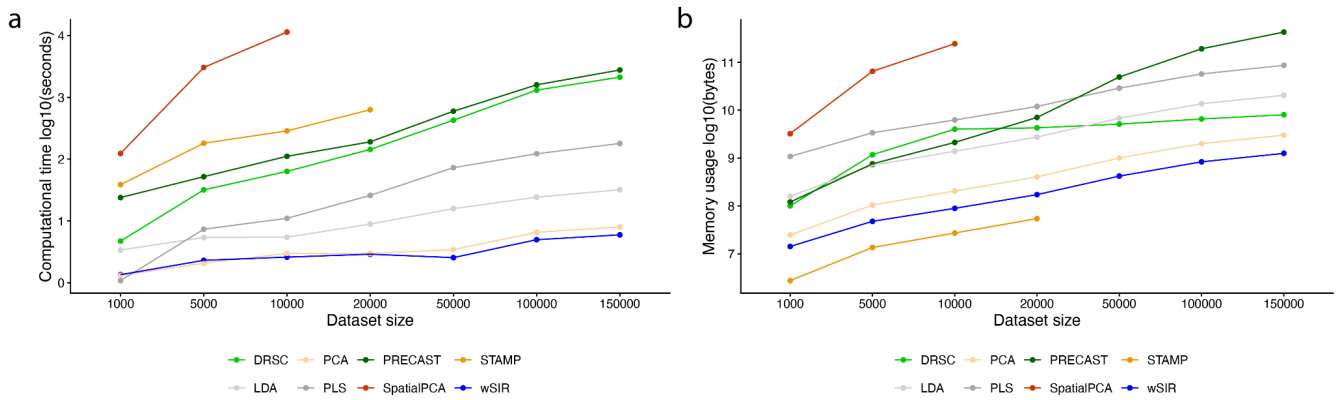
Supplementary Figure 7. a. Barplots displaying bias-corrected distance correlation for three samples (facets) from simulated dataset, comparing wSIR against other spatial dimension reduction methods. b. Line plots displaying neighbourhood similarity for each sample over a range of neighbourhood sizes (x-axis). c. Line plots displaying pseudo-R^2 metric for each sample over different numbers of dimensions for low-dimensional embedding (x-axis). d. UMAP visualisation of wSIR and other spatial dimension reduction embeddings, DR.SC, PRECAST, SpatialPCA, and STAMP (rows) for each sample (columns), coloured by true spatial domain. e. Plots of spatial coordinates of each sample coloured by true spatial domain.

We have added the following text to the manuscript:

Since tissue domain detection is a key analysis task in exploring spatial omics data, we examined wSIR's ability to produce meaningful and stable spatial domains compared to other state-of-the-art methods. While wSIR is fundamentally an approach to embed observations into a reduced dimension space, it can be coupled with unsupervised learning to generate spatially-informed clusters or domains. To facilitate this comparison, we used one real dataset (DLPFC, Methods), and one synthetic simulation dataset (as per Liu et al. (2023a), Methods) that contained what can be considered 'ground truth' domain labels. When we compared wSIR's derived domains against that of other methods including DR.SC, PRECAST, SpatialPCA, and STAMP for the real DLPFC dataset, we observed comparable performance in terms of pseudo-R2 and ARI (Supplementary Figure 6a-d). Similarly, for the synthetic data, we observed both comparable performance in terms of BCDC, neighbourhood similarity and pseudo-R2 (Supplementary Figure 7a-c), as well as observing a consistent visual recapitulation of the striped spatial patterns in this dataset (Supplementary Figure 7d-e). Overall, these analyses demonstrate wSIR's ability to facilitate downstream detection of discrete spatially relevant domains.

3. (Running time and memory usage comparison). While wSIR is claimed to be computationally efficient, the current evaluation lacks systematic benchmarking of runtime and memory usage across datasets.

Thank you for pointing out this omission. To address this, we performed wSIR and other dimension reduction methods on datasets of different sizes, and recorded the computational time and memory cost using the `mark()` function from the bench R package. The results are displayed in the new Supplementary Figure 1 (copied below), which show that wSIR outperforms competing methods both in terms of runtime and memory usage. The discussion of the computational time is also made clear in Section 2.1.



Supplementary Figure 1. a. Benchmarking of computational time (log-scaled y-axis) of methods with varying dataset sizes (x-axis). SpatialPCA and STAMP failed to run within 1 hour for datasets above 10,000 and 20,000 samples respectively. b. As in panel a for memory usage in bytes (log-scaled).

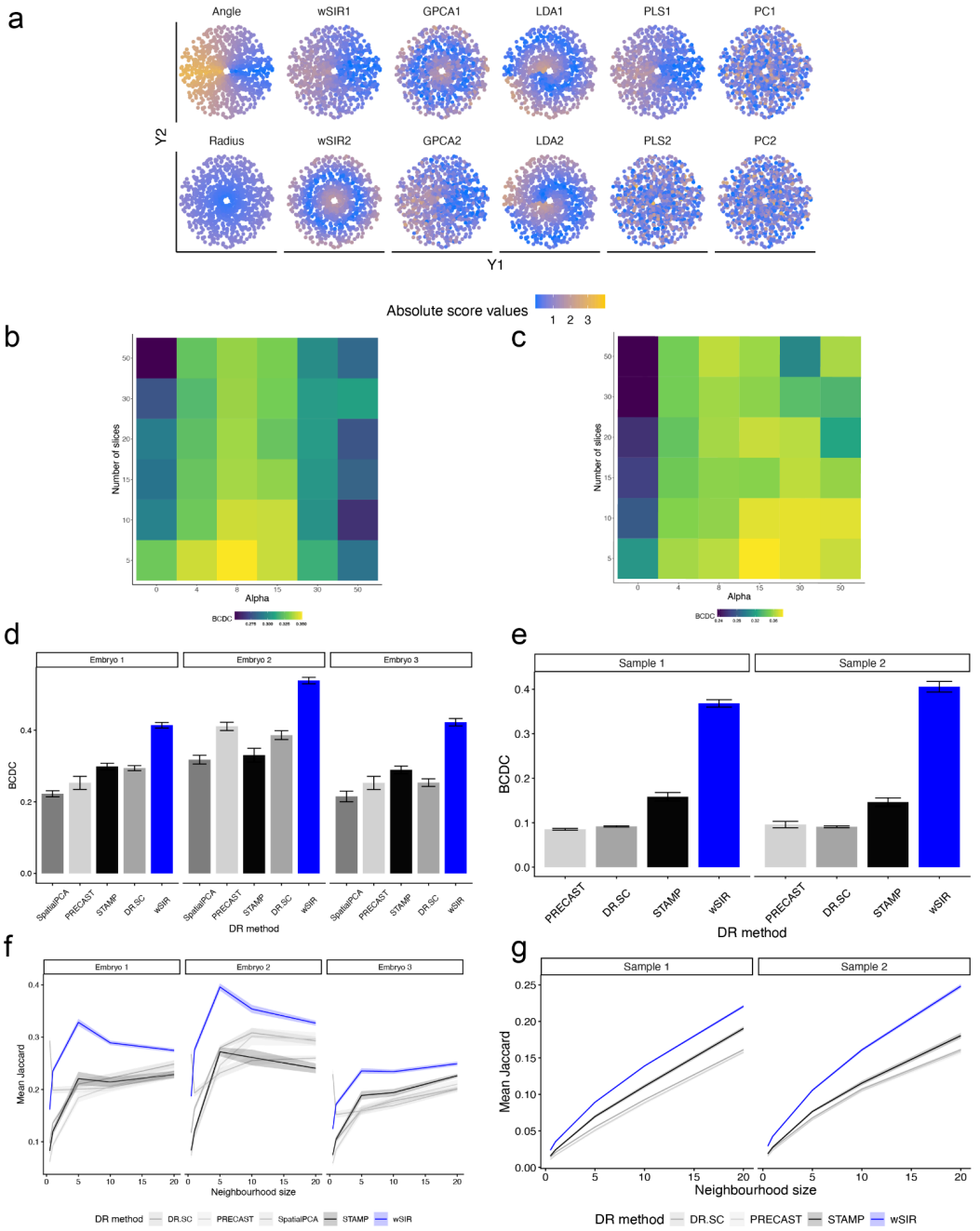
4. (Method comparison). As a dimensionality reduction method specifically developed for spatial transcriptomics, wSIR should be systematically benchmarked against spatially-aware methods (e.g., SpatialPCA, DR-SC, PRECAST, and STAMP) rather than predominantly spatially-unaware methods (e.g., PCA and LDA). This is especially critical for the real-data analyses, where GraphPCA is the only spatially-aware baseline considered.

Thank you for this important point. Our initial reasoning for omitting these other approaches was due to the inability of such methods to be able to project new non-spatial data into the low-dimensional embeddings, however, this does not preclude us from examining 'in-sample' embeddings as suggested here. To address this concern, we have now benchmarked wSIR against the abovementioned spatially-aware dimension reduction methods SpatialPCA, DR-SC, PRECAST, and STAMP on the mouse embryo and the Xenium breast cancer datasets. For each dataset, we evaluated performance using the bias-corrected distance correlation and the neighbourhood similarity metrics. In the revised Figure 2 (copied below), error bars and clouds are based on bootstrap samples of the datasets. We found it is unclear how to use these spatially-aware dimension reduction methods like SpatialPCA, PRECAST, DR.SC and STAMP to project new non-spatial data

into their low-dimensional embeddings, so these methods can't be evaluated in an out-of-sample (e.g cross-validated) study where we need to project non-spatial gene expression without spatial coordinates. Hence, the metrics we reported in Figure 2 are in-sample metrics.

To demonstrate the out-of-sample performance of wSIR, we still keep our comparison of wSIR with non-spatial dimension reduction methods (such as PLS, LDA, and PCA) but moved the result Supplementary Figure 3 (this was originally Figure 2d-g), which demonstrated a better performance of wSIR as well.

We have added a brief description for Spatial PCA, DR-SC, PRECAST and STAMP in Section 4.3, and updated Section 2.6 for the new comparison between wSIR and these methods.



Revised Figure 2(d-g). Bias-corrected distance correlation and neighbourhood similarity metrics for the mouse embryo and Xenium breast cancer datasets using spatially-aware dimension reduction methods. SpatialPCA was not included for the Xenium breast cancer dataset since it did not run within a reasonable time.

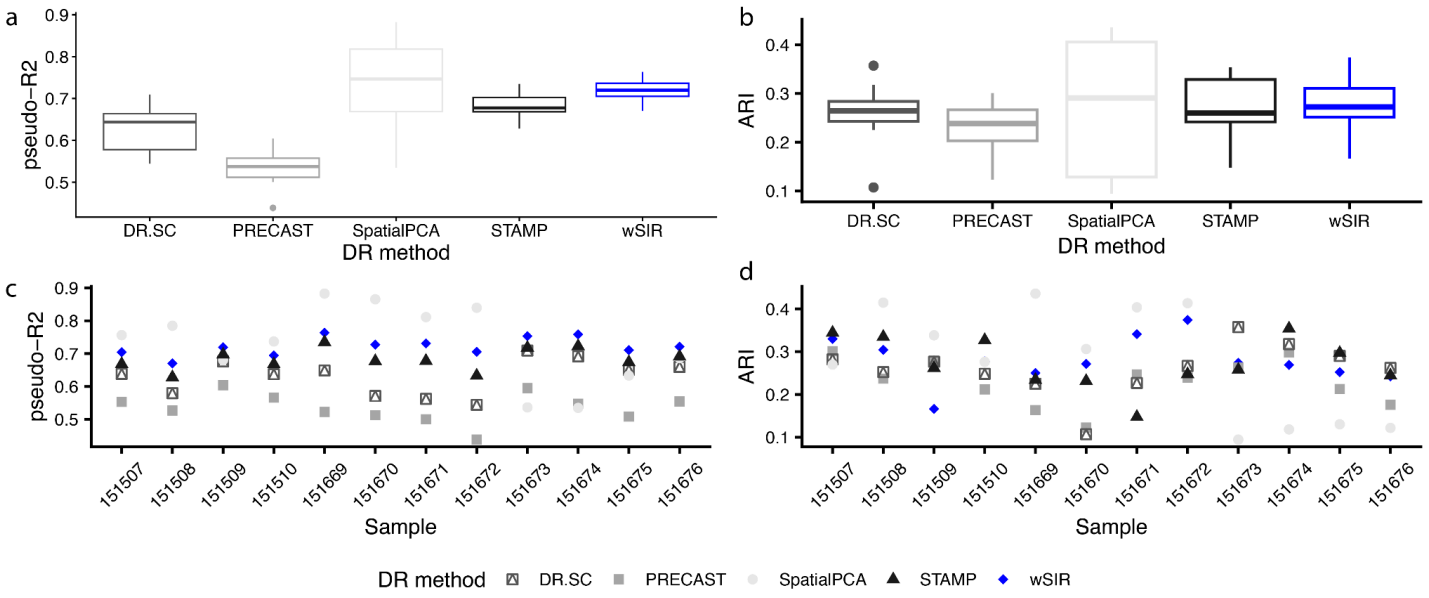
5. (Evaluation metrics). While BCDC measures spatial concordance, it does not directly evaluate the discriminative power of the embeddings for cell type or domain prediction. Incorporating additional metrics like the McFadden-adjusted pseudo-R^2 (as used in SpatialPCA) would provide a more rigorous assessment of the

embeddings' predictive capacity for biological labels. Furthermore, the manuscript should systematically assess the biological relevance of the low-dimensional embeddings by benchmarking performance on critical analytical tasks such as spatial domain detection/cell type clustering using metrics like the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

Thanks for raising this point. It is important to note that the BCDC metric does not rely on any 'ground truth' label, but indeed, it does correspond to concordance between the spatial coordinates and the embedding, which may or may not be biologically driven.

For the suggested additional metrics, we note that one needs a ground truth to calculate them and perform comparison. Hence, in the absence of such labels for our existing mouse embryo and Xenium breast cancer datasets, we have introduced an additional benchmarking dataset to the revised manuscript. We have now added the dorso-lateral pre-frontal cortex dataset[9] that includes annotated spatial domains. For this dataset, we first computed low-dimensional embeddings on each of the twelve samples, then calculated first the pseudo-R^2 treating the spatial domains as the response, and second the ARI when performing clustering on the low-dimensional embeddings. The other spatially-aware dimension reduction techniques (PRECAST, DR.SC, STAMP, SpatialPCA) all include clustering as part of their methods, so we can use those methods to directly obtain clusters. wSIR does not inherently include any clustering component, so we perform k-means clustering on the wSIR embedding to estimate the wSIR clusters and subsequently calculate the ARI. We specify the number of domains in each sample to be the k value in the k-means algorithm.

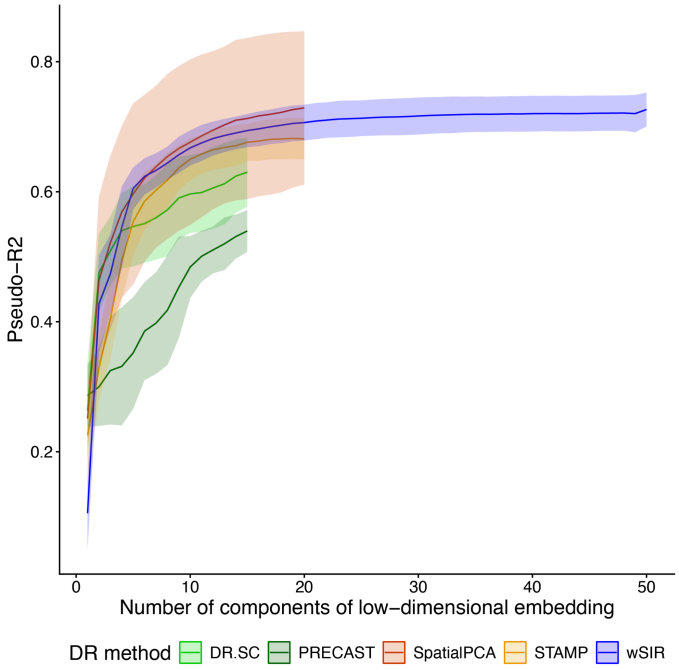
Supplementary Figure 6 (copied below) shows boxplots of the pseudo-R^2 and ARI values from each method in panels a and b, where the boxplots are over the 12 samples. Panels c and d display the pseudo-R^2 and ARI for each of the samples from each method. We observe that although SpatialPCA appears to have the highest pseudo-R^2 and ARI values, wSIR is comparable to existing methods at this task for which it was not specifically created, unlike other approaches such as PRECAST and DR.SC. The comparable performance of wSIR to the other methods suggests that because wSIR can create an embedding that preserves the spatial information in the gene expression matrix, it also performs well on similar tasks, such as clustering spatial domains.



Supplementary Figure 6. a. Boxplot of pseudo-R^2 values for spatial domain detection methods across 12 DLPFC samples. b. As in panel a for ARI values. c. Dotplot of pseudo-R^2 values arranged by each of the 12 DPLFC samples. d. As in panel c for ARI values.

There is a possibility that embeddings with higher dimensions may result in larger pseudo-R^2 values due to the monotonicity of R^2 with added explanatory variables in linear models. In Response Figure R1, we observe that

wSIR has a comparable pseudo-R^2 value compared to other methods when using all components of the embeddings, wSIR is still comparable when only using a smaller subset of the wSIR embedding.

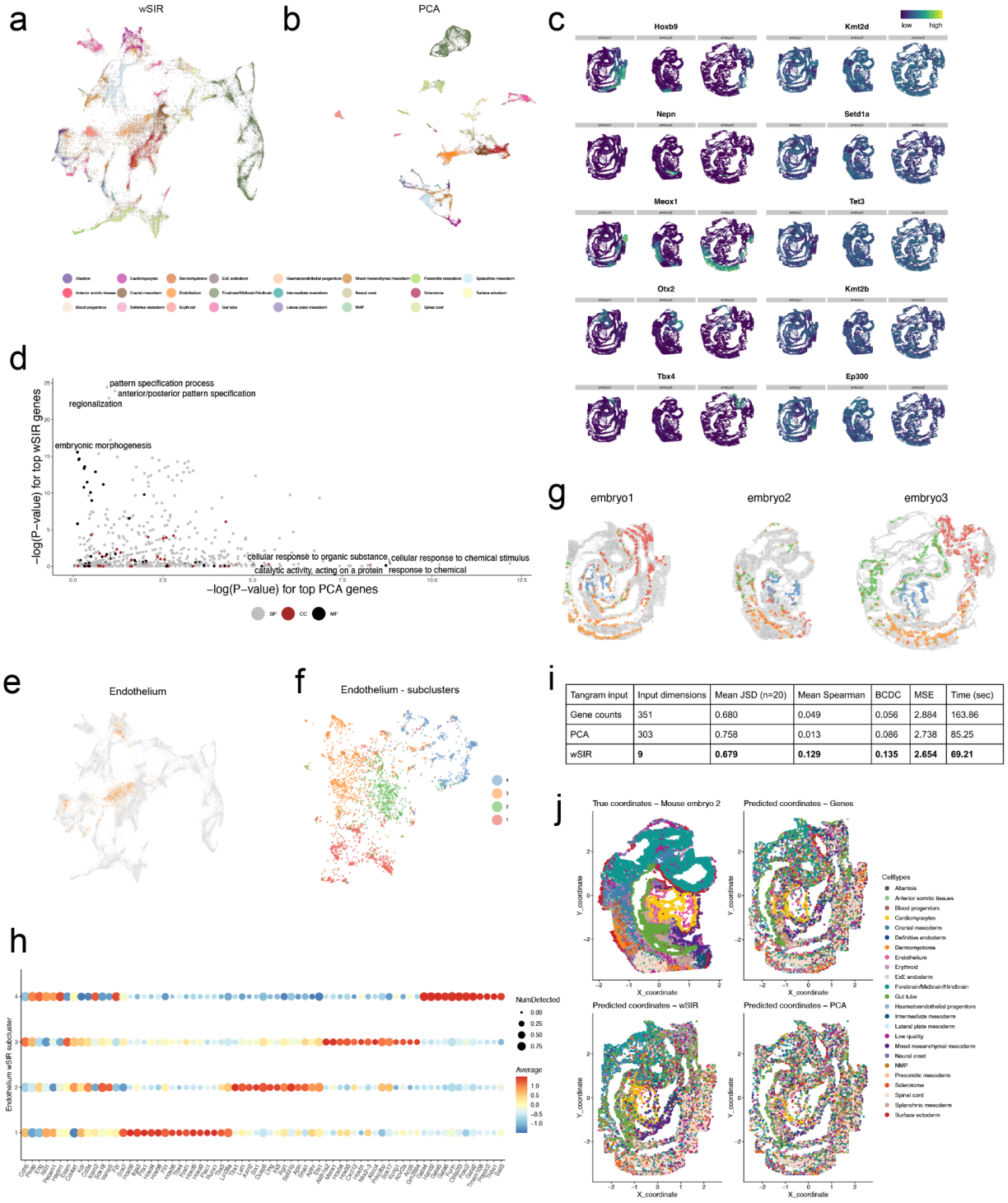


Response Figure R1. Line plot displaying pseudo-R^2 values for different low-dimensional embedding sizes for each spatially-aware dimension reduction method. Lines are the means and ribbon widths represent standard deviations over the 12 DLPFC samples.

We have included a new Section 6.3 to introduce Pseudo-R^2 and ARI performance metrics, and revised Section 2.6 to include a discussion around the results in Supplementary Figure 6.

6. (Spatial prediction evaluation). The use of wSIR embeddings in Tangram demonstrates acceleration in spatial mapping, but spatial reconstruction accuracy should also be quantitatively assessed. Incorporating measurements such as Jensen-Shannon divergence and Spearman's rank correlation (as implemented in CellContrast) would provide a more comprehensive and interpretable evaluation.

Thanks for this suggestion. We have now computed the Jensen-Shannon divergence and the Spearman rank correlation on the Tangram embeddings. The revised Figure 3 (copied below) illustrates the mean Jensen-Shannon divergence (JSD), Spearman correlation, BCDC, and MSE between the true spatial coordinates and the predicted spatial coordinates from each method.



Revised Figure 3. Exploration of mouse development dataset and spatial mapping benchmark with Tangram. a, UMAP of seqFISH mouse development cells using wSIR, coloured by original cell type annotation. b, Same as panel a but using PCA as underlying dimension reduction approach. c, Spatial expression plots of top 5 genes according to wSIR (left column) and PCA (right column). d, Scatterplot displaying GO terms according to -log(P-value) for enrichment of top-ranked genes in PCA (x-axis) and in wSIR (y-axis), where the most significant terms are labelled. e, UMAP as in panel a, highlighting cells belonging to Endothelium. f, UMAP of Endothelium cells using wSIR, coloured by clustering of these cells. g, Spatial plots of Endothelium cells coloured by clusters. h, Expression dotplot of Endothelium clusters, the first 15 genes correspond to markers of Endothelium, and other genes correspond to specific markers for the clusters. The size of each dot represents the proportion of detected expression, and the colour indicates the average expression level relative to each group. i, Tangram benchmarking with the mouse development dataset. The table compares input dimensions, mean JSD for 20 neighbors, bias-corrected distance correlation (BCDC), mean-squared error (MSE), and running time for Tangram with three inputs, including raw gene counts, PCA loadings, and wSIR loadings. j, Spatial plots of true cell types (top left) and Tangram predicted cell types for gene counts (top right), PCA (bottom left) and wSIR (bottom right).

Minor comments:

1. (Introduction, Line 58) MUSTARD is a trajectory-guided dimensionality reduction method for scRNA-seq data, not specifically designed for spatial transcriptomics data.

Thank you for noting this, we have removed this incorrect reference.

2. (Results) For the mouse embryo analysis, the authors should report post-QC cell numbers for all three replicates and include spatial plots of ground truth and annotation results by each method.

Thank you for this, we have added these numbers and included the ground truth annotation in Figure 3j.

3. (Figure 3) Please increase font size in Figure 3. The current text is too small to read clearly.

Thanks, we have fixed the font sizes for readability.

4. (Figure 3h) The legend for dot sizes is partially truncated.

Thanks, we have fixed this error.

5. (Figure 3i) The x-axis label 'WSIR' should be 'wSIR'; The x-axis title 'DR method' should be 'Input type of Tangram'.

We now replaced Figure 3i in the previous version of manuscript to a new Figure 3i that showed more evaluation metrics and displayed the predicted versus actual spatial coordinates.

6. (Evaluation metrics) The methodological details of mean-squared error (MSE) require explicit documentation. Besides, the authors should provide spatial visualizations of reconstructed locations.

Thank you for noting this, we have added the explicit formulation of MSE in Section 6.4. In addition, we have provided the reconstructed locations in the revised Figure 3j, see our response to your comment 6 above.

7. (Format) Adjust paragraph format following equations. When an equation concludes (e.g., Lines 360 and 371), the subsequent paragraph should begin with an indentation.

Thanks, we have fixed this error.

Reviewer: 4
Comments to the author:

The authors develop a modification of SIR to better retain the predictive power for the spatial coordinates. More precisely, the slicing step treats the spatial coordinates as multiple responses, and notably, correlations among slices are incorporated into the kernel matrix M. Both simulation results and case studies indicate that the method yields promising performance across various scenarios, and outperforms existing alternatives. Overall, the manuscript is well written, and the methodology is clearly described and easy to follow.

Thank you for your positive comments on our work.

While the proposed method appears well-motivated and demonstrates promising performance in the current manuscript, a few concerns remain regarding its theoretical foundation and potential limitations. Some issues are listed below:

* Theoretical foundation: The manuscript would benefit from a clearer discussion or formal analysis of the method's theoretical properties, such as the consistency of the estimators. Even partial theoretical results or heuristic arguments could help clarify the method's behavior.

Thank you for your comments. We see theoretical discussion of the method's properties to be out of the scope of this paper. Rather, we suggest that wSIR is a useful tool for computational biology, and we demonstrate this via our results. While this is strictly not related to the new wSIR methodology, we have added more references to the literature regarding the statistical properties of SIR[10–12], and note that exploration of these in the context of wSIR is left for future work in the Discussion.

* Validity of surrogate dependence structure: The use of distance correlation to represent the underlying spatial structure is a reasonable choice, but it would be helpful if more justification for this surrogate could be provided.

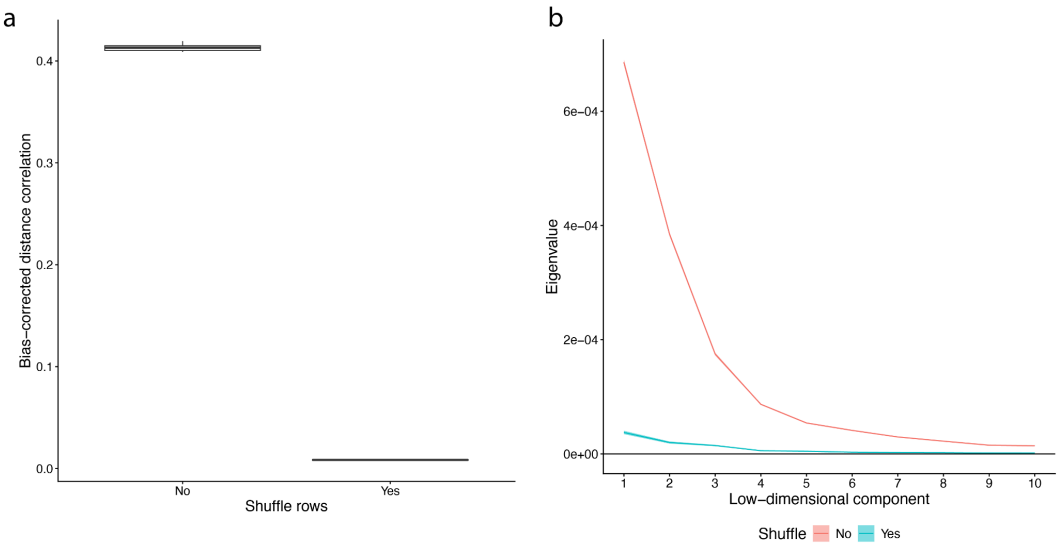
Thank you for raising this point. Our motivation for the use of bias-corrected distance correlation is in the potential issue when comparing distances across different dimensions. For example, when we are comparing the embedding from wSIR to the spatial locations, we may be comparing 30 dimensions to 10 dimensions, which could result in some bias if we used only the typical formula for distance correlation[13]. Even so, it is important to use a variety of evaluation metrics, and we include additional measures such as Jaccard Index to measure the similarity between the neighbourhoods of different sizes in the true spatial coordinates and in the embeddings (Figures 2f-g) in the original manuscript. In response to Reviewer 3, we also provided results for additional metrics using the pseudo-R^2 and Adjusted Rand Index (ARI) metrics, which measure how well the embedding distinguishes spatial domains. As such, we believe we have demonstrated with a variety of quantitative metrics that wSIR is a useful and performant method for spatial omics dimensionality reduction.

* Risk of false discoveries: Although the method seems effective at detecting meaningful spatial patterns in the presented examples, it would be important to evaluate its behavior under null scenarios. For example, when the features are unrelated to the spatial coordinates, would the method still produce spurious signals?

This is a great point. In order to evaluate this, we perform an experiment where we remove the spatial relationship between the gene expression matrix and the spatial coordinates, so we can see how wSIR performs under situations where the spatial coordinates are unrelated to the gene expression features. Specifically, we take one sample of the Mouse Embryo dataset and randomly shuffle the order of the rows so that any relationship between the spatial coordinates and gene expression is gone. We compute the bias-corrected distance correlation between the wSIR embedding and the spatial coordinates to see if wSIR can create a spatially-relevant embedding when there is no spatial signal in the gene expression matrix. We observe that the bias-corrected distance correlation is close to zero when the rows are shuffled, compared to a much larger value when the rows are not shuffled (Response Figure R2). We also compute the eigenvalue of each low-dimensional component when the rows are randomly shuffled or left in their original state. The

eigenvalues can be taken to represent the importance of each feature, so a higher value means that feature is more important for the spatially-informed embedding. Below we plot the first 10 eigenvalues of wSIR performed on the shuffled or original data, and we see that the eigenvalues are much further from zero for the original data than for the shuffled data. This wSIR is not identifying spurious spatial relationships when there is actually no spatial relationship present in the data. In both plots, the boxplots represent bootstrap selection of the dataset.

It should be noted, however, that wSIR will always produce an embedding when run on a valid input dataset even if there is no spatial relationship between the gene expression matrix and the spatial coordinates. It is out of scope of the current work to identify if the features or low-dimensional components are spatially significant, as spatially-variable gene detection methods may aim for. This is a potential direction for future work involving wSIR.



Response Figure R2. A. Boxplot of bias-corrected distance correlation when rows are randomly shuffled compared to no shuffling. B: line plot of eigenvalues of first 10 wSIR directions when gene expression vectors are randomly shuffled between spatial points compared to no shuffling.

* Clarification of limitations: It is helpful to discuss the limitations of the proposed method, such as scenarios where it might seem applicable but is actually unsuitable. This could guide practitioners in avoiding potential misuse.

Thank you for raising this important point. We believe that wSIR is helpful in uncovering spatially relevant variation, but may be limited in identifying non-spatial but still biologically relevant variation. As such, wSIR should be considered as one of many tools that analysts can use for understanding their complex and high dimensional spatial omics data. We have added the following text to the manuscript to illustrate this point, copied below:

> *We envisage wSIR being used in a variety of contexts, especially as spatial transcriptomics data becomes more widespread. Effectively using spatial information is a challenge in contemporary spatial transcriptomics analyses, and simply using tools designed for scRNA-seq can hinder the ability to draw insight in terms of spatial biology. We believe that wSIR is helpful in uncovering spatially relevant variation, but may be limited in identifying non-spatial but still biologically relevant variation. As such, wSIR should be considered as one of many tools that analysts can use for understanding their complex and high dimensional spatial omics data. Overall, we see the potential of wSIR to ensure that spatially relevant variation is not lost, and to enable more comprehensive understanding of spatial biology in several contexts.*

References

1. Yayon, N. *et al.* A spatial human thymus cell atlas mapped to a continuous tissue axis. *Nature* **635**, 708–718 (2024).
2. Li, K.-C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316 (1991).
3. Lin, Q., Zhao, Z. & Liu, J. S. Sparse sliced inverse regression via Lasso. *J. Am. Stat. Assoc.* **114**, 1726–1739 (2019).
4. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2023).
5. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
6. Weisberg, S. Dimension Reduction Regression inR. *J. Stat. Softw.* **7**, (2002).
7. Zhao, Z., Lin, Q. & Liu, J. LassoSIR: Sparsed Sliced Inverse Regression via Lasso. *CRAN: Contributed Packages* The R Foundation https://doi.org/10.32614/cran.package.lassosir (2017).
8. sliced. *PyPI* https://pypi.org/project/sliced/.
9. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
10. Hsing, T. & Carroll, R. J. An asymptotic theory for sliced inverse regression. *Ann. Stat.* **20**, 1040–1061 (1992).
11. Lin, Q., Zhao, Z. & Liu, J. S. On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Stat.* **46**, 580–610 (2018).
12. Lin, Q., Li, X., Huang, D. & Liu, J. S. On the optimality of sliced inverse regression in high dimensions. *Ann. Stat.* **49**, (2021).
13. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *arXiv [math.ST]* (2008).

RESPONSE TO REVIEWERS

We sincerely thank the anonymous reviewers and editor for their second round of helpful feedback. From this, we have revised the manuscript and included a point-by-point response to each reviewer's comments.

In this document, black text represents original reviewer comments, blue text represents author responses. In the revised manuscript, the major changes are highlighted in magenta.

Reviewer #1 (Remarks to the Author):

I appreciate the authors' efforts. All my concerns have now been addressed.

Thank you for your efforts to read our revisions.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

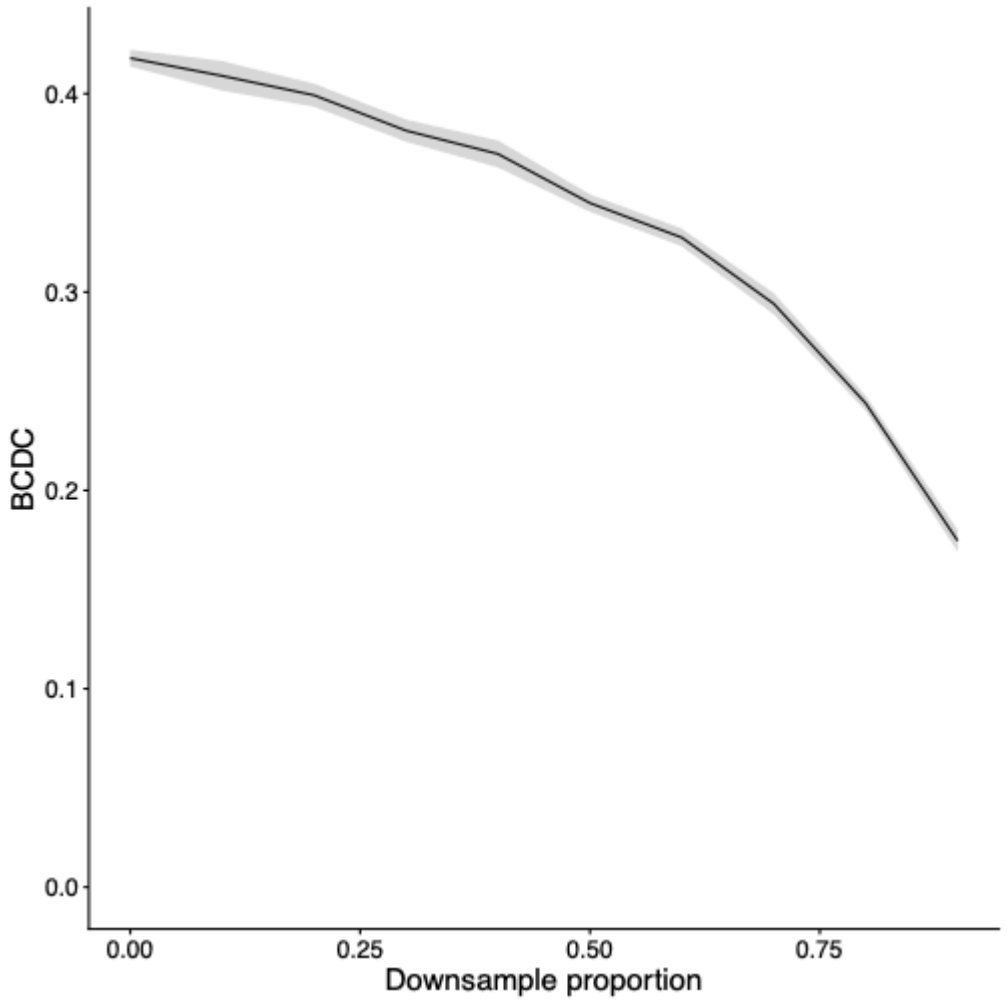
Thank you for your work in co-reviewing our manuscript.

Reviewer #3 (Remarks to the Author):

I appreciate the authors' efforts in addressing my previous concerns. However, there are still several specific points require further clarification.

1. While geometric invariance is a valuable mathematical property, evaluating the algorithm's robustness against varying spot sparsity and noise/dropout levels remains critical. Because wSIR relies heavily on summarizing data within spatial slices, its intra-slice covariance estimations are theoretically highly vulnerable to severe sparsity. Therefore, I suggest the authors perform the sparsity/dropout level simulations to assess the method's robustness.

Thank you for providing this clarification. To address this question, we perform an experiment where we simulate increasing sparsity levels in a dataset, and observe how rapidly and by how much the wSIR performance decreases. For this, we take the mouse embryo sample 1 dataset, and simulate sparsity levels of 0 (no change in expression) up to 90%. At each sparsity level, we compute the bias-corrected distance correlation between the wSIR embedding and the true coordinates. We repeat this process via bootstrapping to obtain the error. We observe that wSIR is quite robust to sparsity, as performance only drops by around 15% at 40% sparsity. The figure is copied below for your convenience, and is also now added as panel f in Supplementary Figure 2.



We conjecture this result is due to the key step of the wSIR algorithm being computing the column means inside each tile, which is not sensitive to uniform sparsity across the dataset. This may not hold if the sparsity is concentrated in a certain region.

2. For the DLPFC dataset, the median Adjusted Rand Index (ARI) for wSIR falls below 0.3. More concerning is that the performance of the competing algorithms reported in this study is significantly lower than benchmark results reported by Hu et al.. The authors must carefully check code and hyperparameter settings for these competing methods to ensure a fair and accurate evaluation. Furthermore, the authors should provide exact P-values (e.g., via paired Wilcoxon signed-rank tests) for the comparisons in Supplementary Figure 6a and 6b.

Thank you for your close attention to this. Regarding the performance of wSIR in clustering, we have two comments. The cortical layers in DLPFC exhibit gradual spatial transitions rather than strictly separable clusters, and the annotations are manually annotated (Maynard et al., 2021) and known to be noisy and partially subjective. As a result, even biologically meaningful embeddings may not achieve high ARI. Second, wSIR is designed for spatially aware dimension reduction rather than direct optimisation of clustering agreement. In Supplementary Figure 6, we saw that ARI from wSIR is comparable to other methods, some of which are more designed for clustering. This comparison further suggests the broad biological relevance of wSIR's embedding.

Regarding the discrepancy between our reported ARI values and those in Hu et al. (2024), we understand how this could cause confusion for readers. We have carefully checked our code as well as the codes written by Hu et al. (2024). We do, however, have reasonable explanations for these differences.

- For the PRECAST method, the Hu et al. (2024) implementation includes integration of all samples prior to clustering. In our analysis, we do not integrate all the samples in order to perform a fair comparison across methods, as the other methods (DR.SC, SpatialPCA, STAMP) do not include built-in data integration. This means that we perform clustering on all samples separately (for all methods), which is why we see a lower performance for PRECAST in our results than in the results of Hu et al. (2024)

- For the SpatialPCA method, we followed the options in the SpatialPCA tutorial for the DLPFC dataset (https://lulushang.org/SpatialPCA_Tutorial/DLPFC.html) that was implemented by the authors of the method (Shang & Zhou, 2022), which is different from the implementation by Hu et al.. Particularly, when computing the loadings and scores for SpatialPCA as implemented in the SpatialPCA_EstimateLoading and SpatialPCA_SpatialPCs functions respectively, we followed Shang & Zhou that used a low-rank approximation of the kernel matrix to accelerate the algorithm by setting fast = TRUE. On the other hand, Hu et al. did not use it (fast = FALSE), i.e. they use the full kernel version.

  We found that the implementation by Hu et al. (2024) led to a higher ARI for SpatialPCA. While the full kernel version of SpatialPCA may lead to improved ARI, this comes at a substantially increased computational cost due to the time complexity of kernel operations. As such, the low-rank approximation is also recommended by Shang & Zhou (2022) for a dataset with a sample size, i.e. number of cells/spots larger than 5000 (see Supplementary Note 1 of Shang & Zhou (2022)). Though the DLPFC samples have slightly fewer cells than this threshold (between 4,200-4,800), our comparison still reflects a practical implementation of SpatialPCA for large spatial transcriptomics data.

- We observe DR.SC attains a similar result as in Hu et al, and STAMP is not benchmarked in that paper.

We have updated the manuscript methods section to reflect these details of our implementation of PRECAST and SpatialPCA for this analysis in sections 4.3 (describing SpatialPCA implementation) and 5.5 (describing the DLPFC dataset).

Exact p-values for Supplementary Figure 6 panels a and b are also computed and described in the figure caption. Thank you for that suggestion.

3. The authors emphasize scalability as a primary superiority of wSIR. While Supplementary Figure 1 shows favorable runtime and memory usage, I am not fully convinced of its practical superiority based on the current presentation. Only a small subset of baseline methods was included in the comparison for Figures 2 and 3. In Figure 4, no other spatially-aware baseline methods were tested.

Thank you for this comment. Figure 2 panels d-g were intended to be a comparison of wSIR against spatially-aware methods. Similar comparisons can be found in Supplementary Figure 3. These figures were not combined because the analysis methods were different: the spatially-aware methods do not allow projection of new non-spatial data, while the non-spatial methods could be evaluated with a traditional train-test split. Therefore, to get the best comparison of wSIR against all the competing methods, we did not use any train/test split with the spatially-aware methods (since that is not possible), and we did use a train/test split against the non-spatial methods. Figure 2 panel a did not use spatially-aware methods because it does not contain cell types (which DR.SC and PRECAST assume), and the simulated data is normally distributed rather than counts-based, which SpatialPCA and STAMP require.

The case studies in Figures 3 and 4 are designed to qualitatively understand what wSIR's output represents biologically, where we study the resulting embedding via UMAP, visualising the most influential genes, and performing clustering on the embedding. These analyses are not designed to show wSIR's superiority over competing methods, which we leave to the figures where we quantitatively evaluate wSIR. Rather, these are designed to give a biological insight into what the wSIR embedding can capture in different situations. We believe such analyses are insightful despite not including other methods, and point to SpatialPCA[4] figure 5 and STAMP[5] figure 6 as case studies in papers on similar topics that do not compare their methods against other methods.

In Figures 3i and 3j, the goal is to assess how different input representations influence the performance of Tangram, rather than to provide a comprehensive benchmark of dimensionality reduction methods. To this end, we selected principal component analysis (PCA) as a widely used non-spatial baseline, and wSIR as a spatially informed alternative, allowing us to isolate the role of spatial structure in the learned embeddings. Importantly, both PCA and wSIR are computationally efficient, which ensures a fair comparison when coupled with Tangram, a deep learning-based method with already expensive computational cost. Including more computationally intensive dimensionality reduction methods would introduce additional variability in runtime and tuning, making it difficult to disentangle the effect of the representation from that of the preprocessing pipeline.

4. The reported mean Spearman correlation for wSIR in the Tangram spatial mapping task is only 0.129. Such a low absolute correlation value raises questions about the actual biological utility of the spatially-informed embeddings for reconstruction.

Thank you for raising this concern. While we observe a low Spearman correlation value in this analysis, we believe that this does not detract from the embedding's biological relevance. The first reason for this is that we do not expect this correlation to be high. Tangram alone does not achieve a high value, yet is still biologically meaningful as demonstrated in its paper[6]. It also does not prioritise Spearman correlation as a metric to determine if the output is biologically meaningful. The second reason is that we present Spearman correlation as an evaluation metric alongside many others. Whereas Spearman correlation measures preservation of global structure, Jensen-Shannon Divergence (JSD) measures preservation of local structure, which is arguably more important in this context. This is because many downstream applications, like cell-cell interactions and domain

detection, rely on local structure rather than global structure.

Minor points:
1. The font sizes in Figure 2 and 3 are still too small to read clearly.

Thank you for pointing this out. We have enlarged the text that was previously too small to enhance readability.

Hu, Y. et al. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. Genome Biol 25, 212 (2024).

Reviewer #3 (Remarks on code availability):

I did not check the code, but the performance of the competing algorithms reported in this study seems too low.

Thank you for pointing this out. We believe we have addressed this in our response to your second comment.

Reviewer #4 (Remarks to the Author):

I thank the authors for their detailed response letter and the revision made to the manuscript. While my previous concerns have been addressed to some extent, several points would still benefit from further clarification and more explicit discussion in the revised manuscript. In this report, I therefore provide more specific follow-up questions and suggestions to help clarify the scope, robustness, and practical use of wSIR. My main comments are listed below.

Thank you for your further suggestions here.

1 (Theoretical foundation) I agree with the authors' statement that "wSIR is a useful tool for computational biology". Thus, it is important to clarify the scenarios and underlying assumptions under which wSIR is expected to be useful, and when it may be less reliable. While a full theoretical analysis may be considered out of scope, it would be helpful to provide a more theoretical discussion of when and why wSIR is expected to work well.

Thank you for your suggestion. We conjecture that theoretically, the consistency of wSIR in the estimation of the dimension reduction subspace between the spatial coordinates and gene expression measurements requires the same mild linearity condition as regular SIR, which is held exactly when the gene expression measurements (in log scale) follow an elliptical distribution (Li, 1991) or approximately when the number of genes is large (Diaconis & Freedman, 1984). Whether this consistency would also require a correct specification of the spatial dependence structure among the tiles depends on how we define a (true) spatial model among the tiles' averaged gene expression. For example, if they follow a spatial lag model, correct specification of the spatial covariance matrix is necessary; however, if they follow a regression model with spatially correlated error, consistency would not require correct specification. Such a model is generally challenging to specify, so we avoid doing it

and construct wSIR as an adaptive and efficient tool for capturing possible spatial dependence when applying SIR to spatial transcriptomics data.

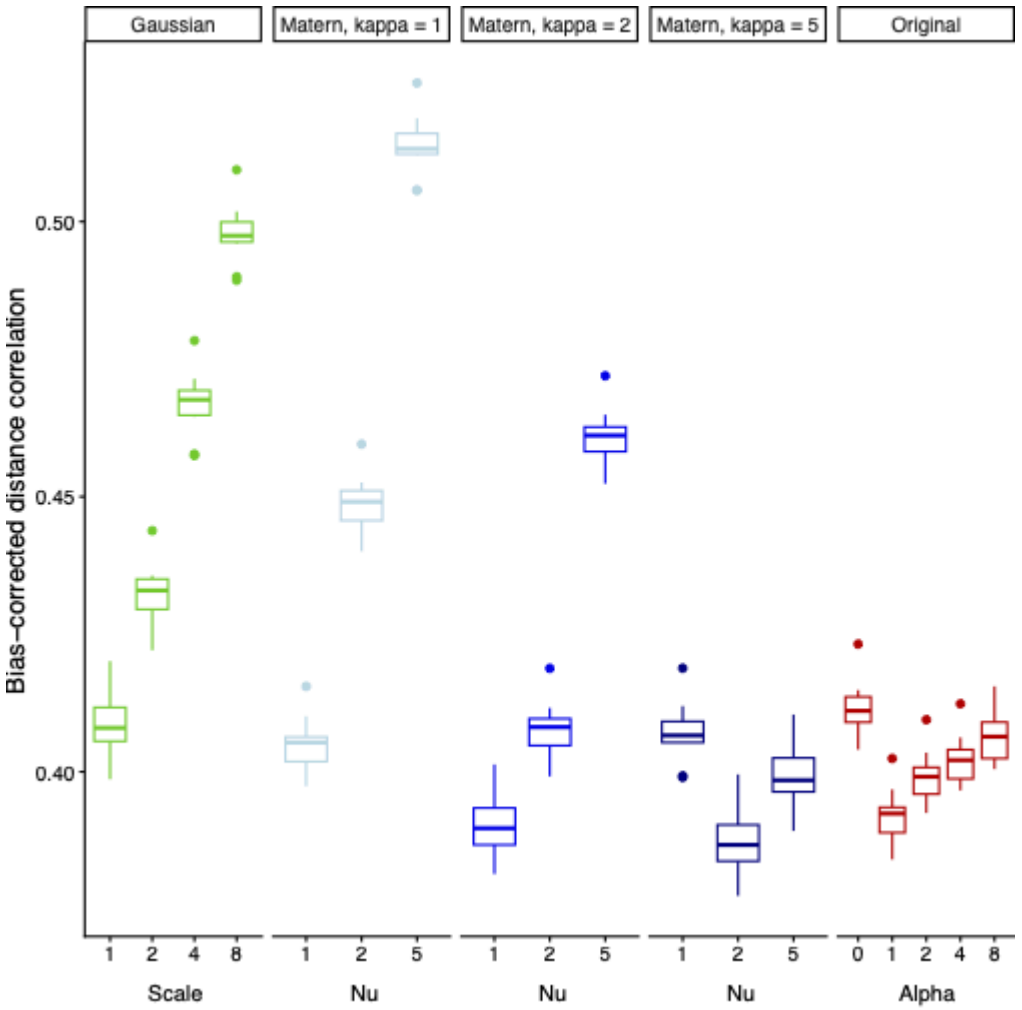
We have modified the Discussion of the revised manuscript to include this point.

2 (Validity of surrogate dependence structure) I appreciate the authors' clarification regarding the motivation for using bias-corrected distance correlation, as well as the inclusion of additional quantitative metrics in the revised manuscript. I would like to clarify that my concern is not about these evaluation metrics, but rather about the working dependence structure used in wSIR,
w_ij = (1-d_ij/d_max)^(1/alpha).

Notably, this form enforces w_ij = 0 for the longest-distance pair(s) with d_ij=d_max, regardless of alpha, which may be restrictive when long-range spatial dependence is present. To strengthen the manuscript, it would be helpful if the authors could either (i) provide a sensitivity analysis under misspecified dependence structures (e.g., exponential or Matern-type correlations with non-negligible long-range dependence), or (ii) discuss a more flexible framework that allows alternative correlation structures, together with practical guidelines for choosing an appropriate structure in applications.

Our choice of weight matrix W to capture spatial dependence among the tiles is motivated by its role in capturing relative spatial proximity rather than specifying a full stochastic covariance model among the tiles. Note that the number of tiles itself is also another hyperparameter that needs to be provided by the user. Hence, our choice w_ij = (1-d_ij/d_max)^(1/alpha) facilitates a dependence structure that provides a scale-free, minimal-free construction that adapts automatically to the spatial configuration and induces effective sparsity. In contrast, Gaussian and Matérn kernels are typically used to model covariance structures of spatial processes, requiring additional parameters (e.g., bandwidth or smoothness) to be estimated. While these kernels offer greater flexibility under explicit probabilistic modeling assumptions, our wSIR method only requires a monotone decay of affinity with distance, for which the proposed form is sufficient and computationally efficient. We added this point to section 4.2 of the revised paper.

Additionally, we compared the performance of wSIR with our original spatial dependence structure and with other spatial dependence structures in the first sample of the mouse embryo dataset. The result is copied below.



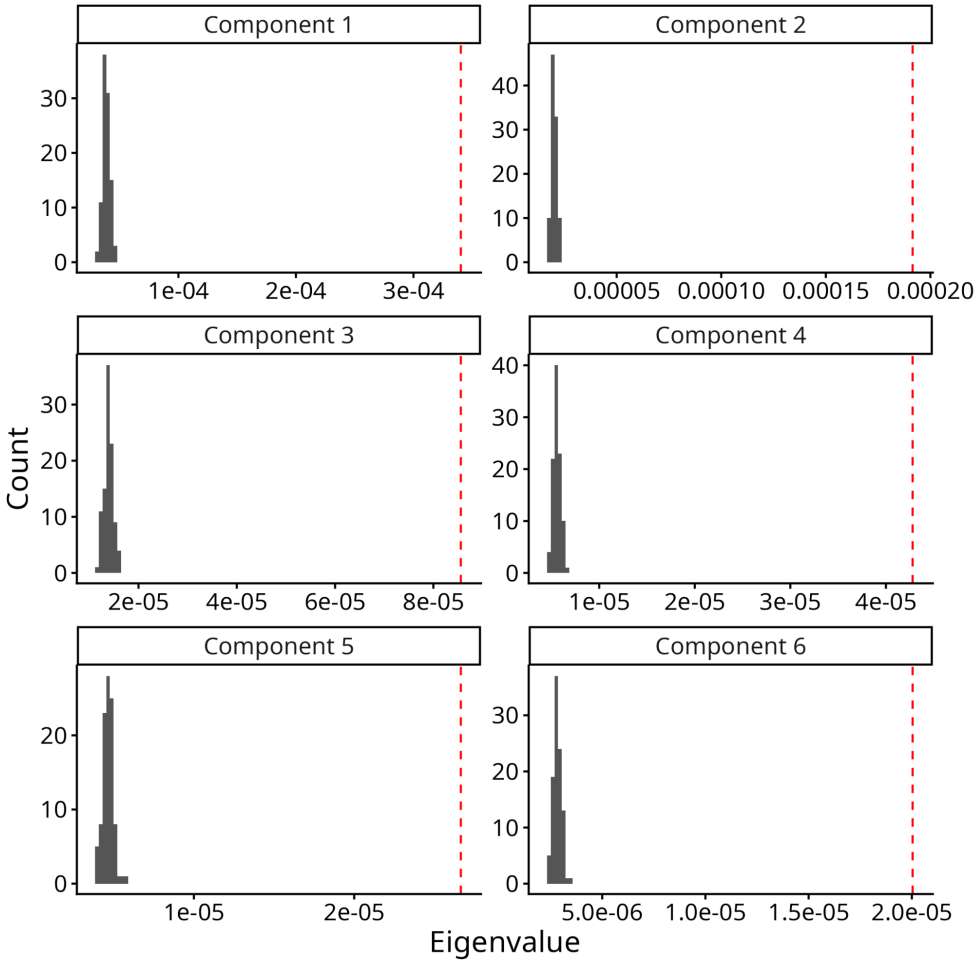
We observe that, with H=8^2 = 64 tiles, wSIR performance is stable across parameter choices for our dependence structure, though the best performance as measured by BCDC is obtained by Gaussian-based or Matern-type kernels for some specific parameters. Nevertheless, the performance of wSIR under these dependence structures is more sensitive to the bandwidth/smoothness. Hence, our dependence structure appears to lead to stable performance for a relatively wide range of alpha. Nevertheless, we now update the Discussion to reflect that the optimal choice of spatial dependence structure using in wSIR can be a worthwhile avenue of future research.

3 (Risk of false discoveries) I appreciate the authors' permutation experiment in the response letter. The reduced leading eigenvalues under this null setting are informative. It would be helpful if the authors could describe this permutation-based procedure more explicitly and provide guidance on how practitioners can use it to assess whether the extracted components reflect real spatial signal.

Thank you for this suggestion. The previous permutation test was performed to compare the wSIR eigenvalues when performed on a dataset with a spatial signal with those corresponding to datasets with no spatial signal. Such an example of the latter can be formed by shuffling the rows of the gene expression matrix X while keeping the same coordinate matrix Y. When applying wSIR on a shuffled gene expression matrix and the

original coordinate Y, we expect that the leading eigenvalues would be much smaller than the leading eigenvalues obtained from applying wSIR to the real (non-shuffled) dataset. Empirically, we can form a large number of shuffled gene expression matrices and obtain the corresponding eigenvalues for each of them; finally, we compute an empirical p-value by the proportion of such eigenvalues that are as large as the one in the real dataset. A small p-value would indicate a stronger spatial signal to the practitioners.

We now added the permutation test to the Supplementary Note A2 of the revised paper, including with the empirical results of this procedure for the first mouse embryo data in the Supplementary Figure 8 (copied below for your convenience)



Supplementary Figure 8. Null distribution for the six largest eigenvalues of wSIR obtained from shuffling the first mouse embryo dataset 100 times. The red dashed line indicates the corresponding eigenvalue on the real (non-shuffled) data

4 (Clarification of limitations) I appreciate the authors' added discussion on the potential limitation that wSIR may be less effective at identifying non-spatial but still biologically relevant variation. In addition to the robustness issues discussed above regarding the working dependence structure, it would be helpful to further clarify the scope of applicability by discussing a few concrete scenarios where the method may be less appropriate. For

example, it would be helpful to comment on whether wSIR remains reliable when some covariates are categorical. If not, it may be useful to briefly discuss whether extensions such as partial dimension reduction (see, e.g., Li, Cook, and Chiaromonte, 2003) could be adapted to this setting.

wSIR is designed specifically for spatial omics datasets that contain only spatial coordinates and gene expression measurements at the single cell/spot level, where the latter are continuous covariates. The core innovation of wSIR is to introduce a weight matrix W that captures spatial dependence among slices in the eigendecomposition step of SIR, leading to a dimension reduction subspace of the gene expressions that captures spatial information. Hence, we expect that any existing extension to SIR can also be adapted to wSIR; for example, with the presence of a categorical covariate, we can also perform a partial wSIR method by first performing wSIR conditionally on each category, then accumulating all the estimated subspaces in a similar manner to section 4 of Li, Cook, and Chiaromonte (2003).

We have also updated the text in the discussion for this future research.

5 (Minor) I noticed that Figure 2 has been updated substantially, but its caption still appears to correspond to the previous version. I would recommend that the authors carefully proofread the revised manuscript to ensure that all elements (figures, captions, etc.) are consistent with the updated content.

Thank you for your careful review and our sincere apologies for this oversight. This has now been corrected.

References:

1) Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414), 316–327. https://doi.org/10.2307/2290563

2) Diaconis, P., & Freedman, D. (1984). Asymptotics of Graphical Projection Pursuit. *The Annals of Statistics*, 12(3), 793–815. http://www.jstor.org/stable/2240961

3) Li, B., Cook, R. D., & Chiaromonte, F. (2003). Dimension Reduction for the Conditional Mean in Regressions with Categorical Predictors. *The Annals of Statistics*, 31(5), 1636–1668. http://www.jstor.org/stable/3448390

4) Shang, L., Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat Commun* 13, 7203 (2022). https://doi.org/10.1038/s41467-022-34879-1

5) Zhong, C., Ang, K.S. & Chen, J. Interpretable spatially aware dimension reduction of spatial transcriptomics with STAMP. *Nat Methods* 21, 2072–2083 (2024). https://doi.org/10.1038/s41592-024-02463-8

6) Biancalani, T., Scalia, G., Buffoni, L. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* 18, 1352–1362 (2021). https://doi.org/10.1038/s41592-021-01264-7

7) Maynard, K.R., Collado-Torres, L., Weber, L.M. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 24, 425–436 (2021). https://doi.org/10.1038/s41593-020-00787-0

RESPONSE TO REVIEWERS

We sincerely thank the anonymous reviewers and editor for their thoughtful and constructive feedback. As a result, we have revised the manuscript and included a point-by-point response to each reviewer's comments.

In this document, *italic black text* represents original reviewer comments, blue text represents our responses.

Reviewer #4

I appreciate the authors' efforts in revising the manuscript. My previous concerns have been largely addressed.

At this stage, I have only one remaining suggestion. The comparison of different dependence structure models discussed in the response to Comment 2 (particularly the figure and the corresponding explanation) is quite informative. I recommend incorporating these results into the manuscript, as they provide more direct and intuitive evidence supporting the advantage of the proposed method.

Other than this, I have no further comments. The manuscript is now in good shape.

Response: Thank you for your suggestion. We have now included that figure to the Supplementary Information - see the new Supplementary Figure 3. We also referred to this figure in the paper in the first paragraph of page 5 when we discuss the robustness of wSIR.