Supplemental Materials for Kernel-Bounded Clustering for spatial transcriptomics enables scalable discovery of complex spatial domains

Hang Zhang[1,2,+], Yi Zhang[1,2,+], Kai Ming Ting[1,2,*], Jie Zhang[1,2,*], and Qiuran Zhao[1,2]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China

[2]School of Artificial Intelligence, Nanjing University, Nanjing, 210023, China

[+]Contributed equally to this work

[*]Correspondence: tingkm@nju.edu.cn, zhangj_ai@nju.edu.cn

ABSTRACT

This document contains four Supplemental Texts:

S1: lists the 'Data availability', 'Code availability', 'Evaluation metrics' and 'Parameter settings'.

S2: Results analyse of HER2 tumor data and parameter sensitivity analyse.

S3: describes a comparison between Gaussian kernel, Adaptive Gaussian kernel and Isolation kernel that can be employed in KBC. Example comparison results are also provided.

S4: provides the KBC clustering outcome of an example single-cell resolution dataset

Supplemental Text S1

Data availability

All original data supporting the findings of the study are publicly available online.

The human dorsolateral prefrontal cortex (DLPFC (Maynard et al. 2021)) datasets were downloaded from spatialLIBD. It contains 12 sections, each with 3000-4000 spots, spanning six neural layers and the white matter. The labels were manually annotated by the authors.

Mouse hippocampus Slide-seq V2 data were downloaded from the Broad Institute Single Cell portal. Following Shang et al. 2022, we used the file "Puck_200115_08" in our study. The dataset contains approximately 23,000 genes and 53,000 spatial locations.

The HER2 breast tumor dataset was downloaded from github: her2st. The dataset contains eight samples with pathologist-annotated labels, and we used the H1 sample to demonstrate the detailed result. The sample sizes are small, containing approximately 100-600 spots.

The Stereo-seq data of mouse olfactory bulb tissue was downloaded from github: SEDR_analyses.

The preprocessed spatial genomics datasets we used can be found at: https://github.com/IsolationKernel.

Code availability

The KBC software code is publicly available at https://github.com/IsolationKernel. The source code is released under a non-commercial use license.

Other methods in the paper are available online, as follows:

- SpatialPCA (https://github.com/shangll123/SpatialPCA),
- Stagate (https://github.com/zhanglabtools/STAGATE),
- BayesSpace (https://github.com/edward130603/BayesSpace),
- SpaGCN (https://github.com/jianhuupenn/SpaGCN),

- stLearn (https://stlearn.readthedocs.io).

We followed their tutorials to preprocess the raw data and used the packages implemented by original authors for each competing method to conduct the experiments.

Evaluation metrics

We have used two commonly used metrics in evaluating each clustering outcome of an algorithm. They are Adjusted Rank Index (ARI) and Normalized Mutual information (NMI), given as follows:

$$\text{ARI}(T,P) = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

and

$$\text{NMI}(T,P) = \frac{\sum_{i=1}^{C_T}\sum_{j=1}^{C_P}\left|T_i \cap P_j\right| \log \frac{n\left|T_i \cap P_j\right|}{|T_i| \times \left|P_j\right|}}{\max\left(-\sum_{i=1}^{C_T}|T_i| \log \frac{|T_i|}{n}, -\sum_{j=1}^{C_P}\left|P_j\right| \log \frac{\left|P_j\right|}{n}\right)},$$

where T denotes the ground-truth labels of data points, and P denotes the clustering labelled outcomes of data points. a is the number of pairs of two points in the same group in both T and P; b is the number of pairs of two points in different groups in both T and P; c is the number of pairs of two points in the same group in P but in different groups in T; and d is the number of pairs of two points in different groups in P but in the same group in T. C_T and C_P are the numbers of clusters in T and P, respectively; and n is the number of points in T or P.

Both ARI and NMI range from 0 to 1, the larger the better.

Indeed, there is no universally good metric to assess the clustering outcome of a clustering algorithm (see Section 6.9 in "Data mining: the textbook" (Aggarwal 2015)). That is why multiple metrics are often used in assessing the clustering performance of different clustering algorithms.

Parameter settings

Table S1 shows parameter settings used in the experiments for all the methods. Each clustering algorithm is configured with the same number of clusters which matches the ground-truth number of clusters in a dataset.

Table S1. Parameter search ranges used in the experiments.

	Parameter search range
Gaussian kernel	$\sigma \in \{2^m \mid m \in \{-5,-4,...,4,5\}\}$
Isolation Kernel	$\psi \in \{16,32,64,128\}, t = 100$
KBC	$\tau \in \{0.05 + 0.1 * a \mid a \in \{0,1,...,9\}\}$
WL	$h = 7$
SpatialPCA	$pc \in \{9,12,15,18,21\}$
Walktrap	$knearest \in \{sqrt(n) - 5, sqrt(n) + 5\}$
Kmeans	$n\_init \in \{5,10,15,20\}$
Mclust	default
SpaGCN	$histology \in \{true, false\}, init \in \{$"$louvain$","$kmeans$"$\}$
Stagate	$rad\_cutoff \in \{50,150,160\}$ default as in tutorial
stlearn	default as in tutorial
BayesSpace	$pc \in \{7,9,12\}; init\_method \in \{$"$mclust$","$kmeans$"$\}$

Table S2. Data transformation methods and clustering methods employed in each of the algorithms used in the experiments

	Data Transformation	Initial Clusters	Clustering
KBC	WL or SpatialPCA	Connected components	KBC
SpatialPCA	SpatialPCA	NA	Walktrap
Stagate	Graph attention auto-encoder	NA	Mclust
SpaGCN	GCN	Louvain or Kmeans	Deep learning clustering
BayesSpace	Bayesian statistical method	Kmeans or Mclust	Bayesian Clustering
StLearn	SpatialMorphological gene Expression (SME) normalization	NA	Kmeans/Louvain

Table S2 shows data transformation and clustering employed in each of the algorithms used in the experiments. The initial clusters of KBC (in step 1 in Algorithm 1) is obtained as connected components by using the standard 'concomp' function in Matlab, where each connect component is an initial cluster.

Table S3 summarizes the graph embedding methods used with KBC, and the datasets in which they have been applied.

Table S3. The graph embedding methods used with KBC in the experiments.

SpatialPCA+KBC	Second Ablation Study: DLPFC Simulated datasets	Fig. 3 & 5 Fig. S1
WL+KBC	HER2 tumor, mouse hippocampus, DLPFC Simulated datasets, mouse olfactory bulb	Fig. 7, 8 & 9 Fig. S1, S2, S3, S4, S5, S6, S7 & S8

Supplemental Text S2

Further analysis of the KBC clustering outcomes of the HER2 tumor data

This section examines two key discrepancies between the ground-truth labels (Figure S1A) and the clusters identified by KBC (Figure S1B). For the first discrepancy, we divided the cells into several sub-groups with labels 1, 2, and 3 (Figure S1C), and conducted differential expression analyses (with seurat package) on regions 1, 2, and 3: comparing three pairs of regions: 2 versus 1, 2 versus 3, and 1 & 2 versus 3. We labeled regions 4, 5, and 6 for the second discrepancy and conducted differential expression analyses comparing three pairs of regions: 5 VS. 6, 4 VS. 6, and 4 VS. 5. In each differential expression analysis, we took the union of all significant genes (adjusted p-value ≤ 0.05 under the Wilcoxon Rank-Sum test) and visualized their expression patterns through heatmaps. They are shown in Figures S1D and S1E, respectively.

Note that, in Figure S1D, there is only one significant gene, TIMP1, detected when comparing the region-pair 1 VS. 2. However, we detected 7 and 35 significant genes when comparing 2 VS. 3, and 1 & 2 VS. 3, respectively; and the example top ranked genes are S100A9, HLA.B & B2M and IL7R, C7orf73 & LCP1, respectively. In short, regions 1 & 2 exhibit similar gene expression patterns, both of which differ significantly from region 3. This is consistent with the clustering outcome of KBC.

In Figure S1E, we detected 11, 6, and 4 differentially expressed genes when comparing three pairs of regions: 5 VS. 6, 4 VS. 6, and 4 VS. 5, respectively. Examples of top ranked genes are IGHA1, PFN1, APOD for the first pair; SCGB2A2, CFD, APOD for the second pair; and PFN1, PIEZO1, IGHA1 for the third pair. Note that region 6 has been identified by KBC (in Figure S1B) as a cluster differs from region 4 but belongs to the same cluster as region 5. In contrast, the ground-truth labels shown in Figure S1A indicate that region 6 belongs to the same group as region 4 but differs from region 5. The significant differences in genes among these three regions indicate that there are ambiguities in clustering, partially explaining the second discrepancy.

The above two discrepancies between KBC's clustering outcomes and the ground-truth labels provide a ground for further examination to ascertain whether there is a mistake in data collection or human labeling.

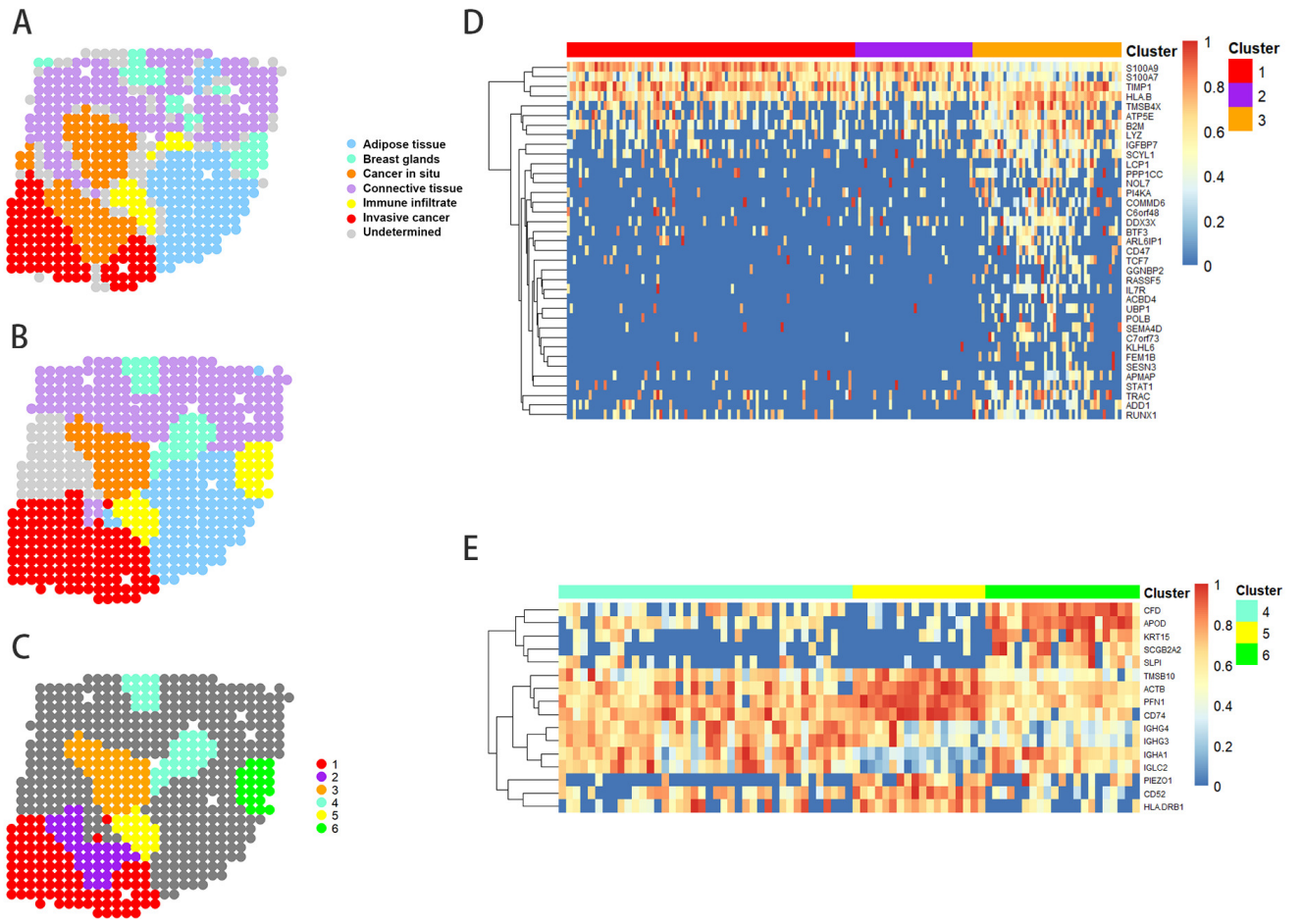

Figure S1. Further analysis of the KBC clustering outcomes of the HER2 tumor data. A, the ground-truth labels. B, the clusters predicted by the KBC clustering method. C, the sub-groups 1-6 obtained by comparing the ground-truth regions and KBC predicted regions. D, the differentially expressed genes by comparing regions 1, 2 and 3, shown in C. E, the differentially expressed genes by comparing regions 4, 5 and 6, shown in C.

Sensitivity to the number of principal components used

We have used 15 principal components in the data processing thus far. Here we examine the effect of this number on the clustering performance on both the simulated HVG and SVG datasets. Figure S2 shows that WL+KBC produces similar clustering outcomes regardless of the number of principal components selected (between 9 to 21) during preprocessing.

Sensitivity to the parameters ψ and τ in KBC

KBC has two parameters ψ (sample size to build Isolation Kernel) and τ (similarity threshold used in KBC) which shall be tuned for a dataset. Here we investigate their sensitivity on the simulated datasets. Figure S3 shows that

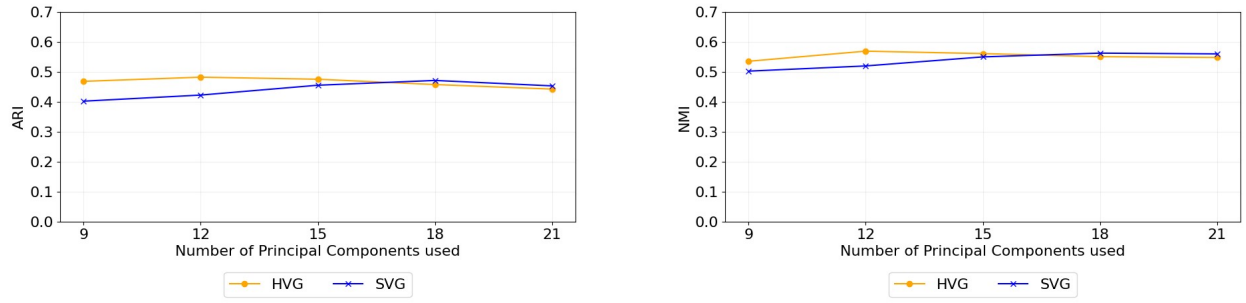

Figure S2. Sensitivity to the number of principal components used in WL+KBC in terms of ARI and NMI on the HVG and SVG simulated datasets. Each point in a plot is based on the median score out of the 12 sections.

KBC is not very sensitive to these parameter settings, especially when $\psi \in [8, 32]$ and $\tau \in [.4, .6]$.

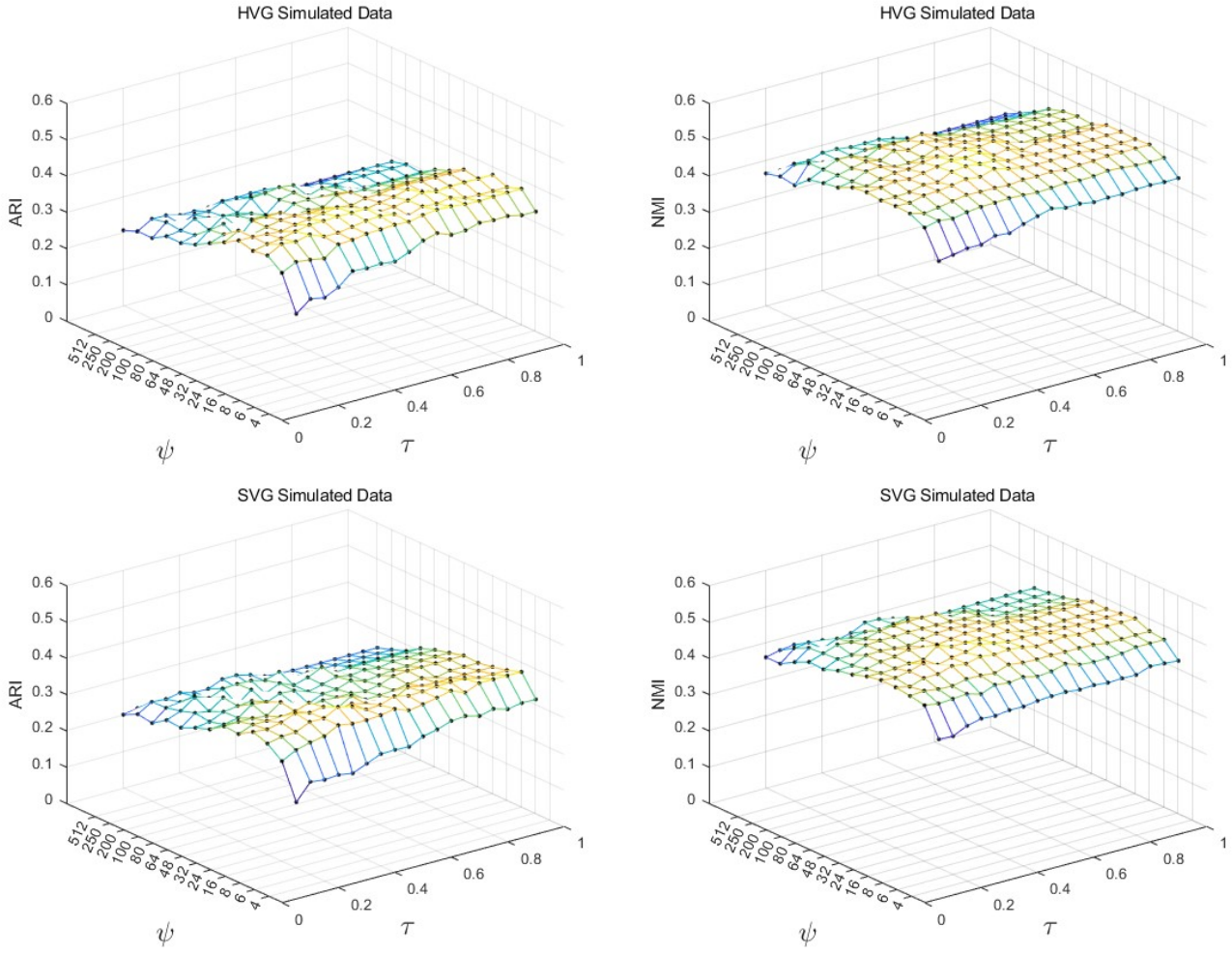

Figure S3. Sensitivity to the KBC parameters ψ and τ in terms of ARI and NMI on the HVG and SVG simulated datasets. Each point in a plot is based on the median score out of the 12 sections.

Sensitivity to the parameter h in WL

We have used $h = 7$ in WL for all experiments. Here we examine the sensitivity of this parameter on the simulated HVG and SVG datasets. Figure S4 shows that WL+KBC produces similar clustering outcomes regardless of the

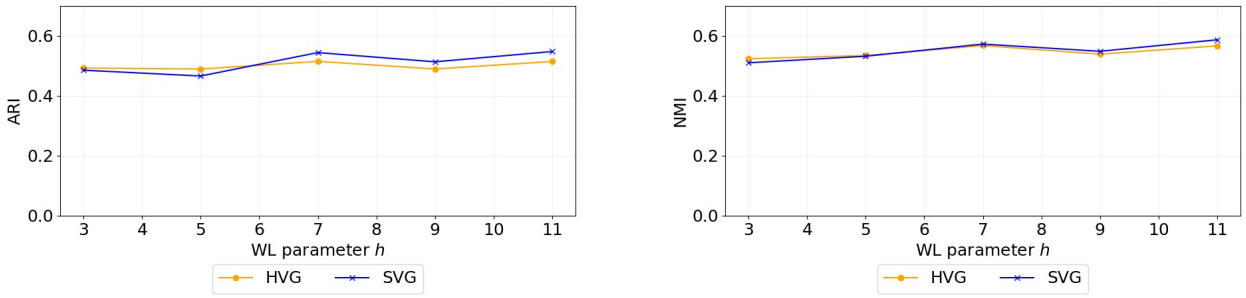

Figure S4. Sensitivity to the number of WL parameter h used in WL+KBC in terms of ARI and NMI on the HVG and SVG simulated datasets (slice 151671). The number of principal components used in WL is set to 15, ψ and τ of KBC are set to 16 and 0.5, respectively.

parameter h setting in between 3 and 11.

Supplemental Text S3

IKBC versus GKBC & AGKBC

We have used a recently introduced Isolation Kernel (Ting et al. 2018) in KBC in this paper. However, KBC admits a commonly used kernel such as Gaussian kernel or Adaptive Gaussian kernel (Zelnik-Manor et al. 2005).

Gaussian Kernel and Adaptive Gaussian Kernel

For any point $x, y \in \mathbb{R}^d$, Gaussian kernel is defined as:

$$\kappa_\sigma(x,y) = exp(\frac{-||x-y||^2}{2\sigma^2}),$$

where σ denotes the bandwidth of Gaussian Kernel.

Note that Isolation Kernel is a data dependent kernel which derives its feature map from a dataset directly, and it has no closed form expression. In contrast, Gaussian kernel, like most other commonly used kernels, is a data independent kernel, which has a closed form expression.

Adaptive Gaussian Kernel (Zelnik-Manor et al. 2005) is defined as follows:

$$\kappa_k(x,y) = exp(\frac{-||x-y||^2}{\sigma_x\sigma_y}),$$

where σ_x is the distance between x and k-th nearest neighbor of x.

We refer to Isolation Kernel-based, Gaussian Kernel-based and Adaptive Gaussian Kernel-based KBCs as IKBC, GKBC and AGKBC, respectively. We use the simulated SVG dataset in the comparison.

Comparison results

The clustering result of the comparison on the simulated SVG dataset (as used in the Supplemental Text S2) is summarized in Fig. S5. It shows that IKBC produces better clustering outcomes than GKBC and AGKBC. AGKBC and GKBC perform comparably (having similar median ARI and NMI). It shows that though AGK is adaptive to

local density, AGK is sensitive to the parameter k setting because it relies on k nearest neighbor distance which prevents it to adapt to a complex distribution with varied densities (see the discussion of this issue in Bandaragoda et al. 2018; Y Zhu et al. 2021).

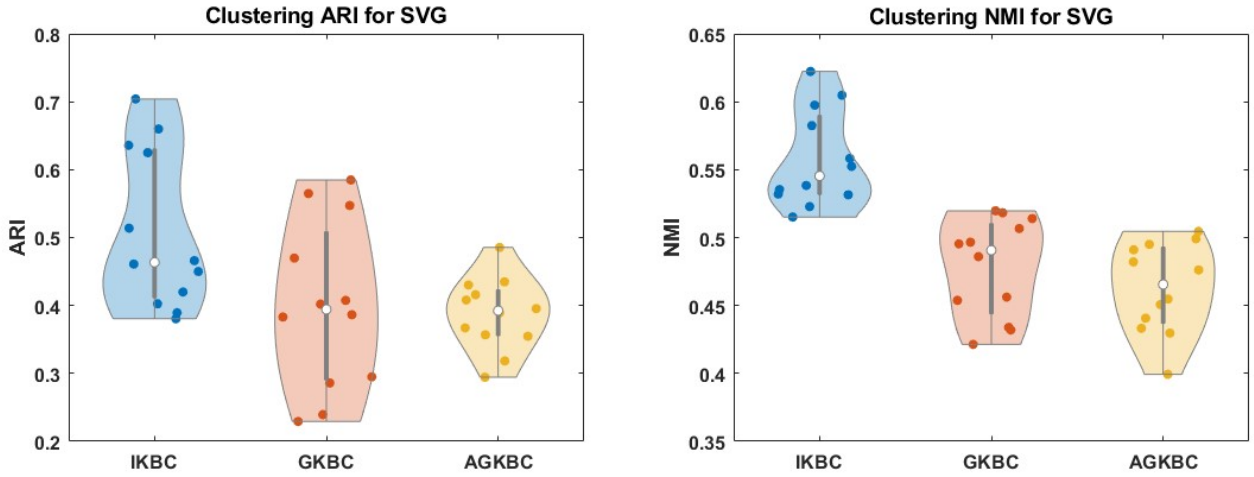

Figure S5. A comparison of IKBC versus GKBC & AGKBC. Violin plots on the simulated SVG dataset for all 12 tissue slices in terms of ARI and NMI. The white point is the median value of the 12 results. All methods use the WL embedding here.

Figure S6 shows an example comparison of GK, AGK, and IK on one slice of the simulated SVG dataset. The first row shows the similarity matrix of each kernel via heatmap, where the (i, j) value in the matrix represents the similarity (as measured by a kernel) between the i-th point and the j-th point.

From the similarity matrix of GK, we can see that the pink cluster has the highest density (because of the highest similarity), while the adjacent yellow cluster has very low density. As a result of the huge difference in density between the two neighboring clusters, GKBC extends the pink cluster to swallow the yellow cluster (shown in the second and third rows). The data-dependent AGK and IK reduce the density difference between the two clusters. This enables AGKBC and IKBC to avoid over-extending the pink cluster. But AGK has made an adverse effect that the blue and yellow (and also the green) clusters become very similar (see the outer box enclosing the three inner boxes in the similarity matrix of AGK), causing AGKBC to largely merge the three clusters. The final outcome is that GKBC and AGKBC have similar ARI/NMI (see the results in the caption). In contrast, the similarity matrix of IK has four clear boxes depicting the four clusters (except the green cluster), enabling IKBC to correctly produce

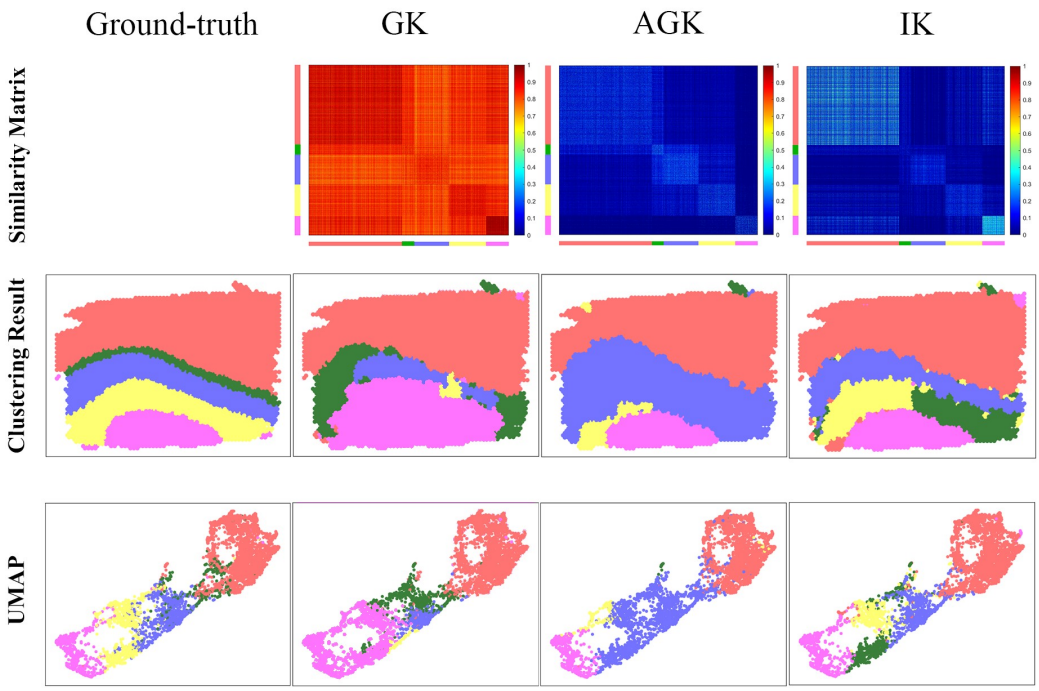

Figure S6. Results of the further ablation studies on three kernels and their associated KBC clustering outcomes in the original space and the reduced 2-D space by UMAP (McInnes et al. 2018) derived from the graph embedded space on the simulated SVGs (tissue slice 151671 of DLPFC dataset). GKBC, AGKBC and IKBC have ARI = 0.68, 0.62 and 0.75, respectively; and NMI = 0.63, 0.62 and 0.71, respectively.

these four clusters (with one caveat—see the next paragraph).

Note that none of the three algorithms could correctly cluster the green cluster because the number of points is very small. This has caused each of three clustering algorithms to split one cluster into two clusters. Compared with the other two methods, IKBC produced better blue, yellow and red clusters, having the highest NMI and ARI among the three methods.

In summary, a data independent kernel, while faithfully reveals the density of each cluster, the resultant clustering algorithm would bias towards high density clusters. This bias often yields a dense cluster to encroach on a neighboring cluster with low density, merging the two clusters as a result. The data dependent AGK attempts to correct this bias by using a single parameter k via a k-nearest-neighbor method. However, in a complex distribution with many clusters of varied densities in the embedded space (as in the example shown in Figure S6), it failed to

adapt well to all clusters. It corrected one (pink) cluster and over-corrected the blue and yellow clusters. The data dependent IK has none of the issues mentioned above.

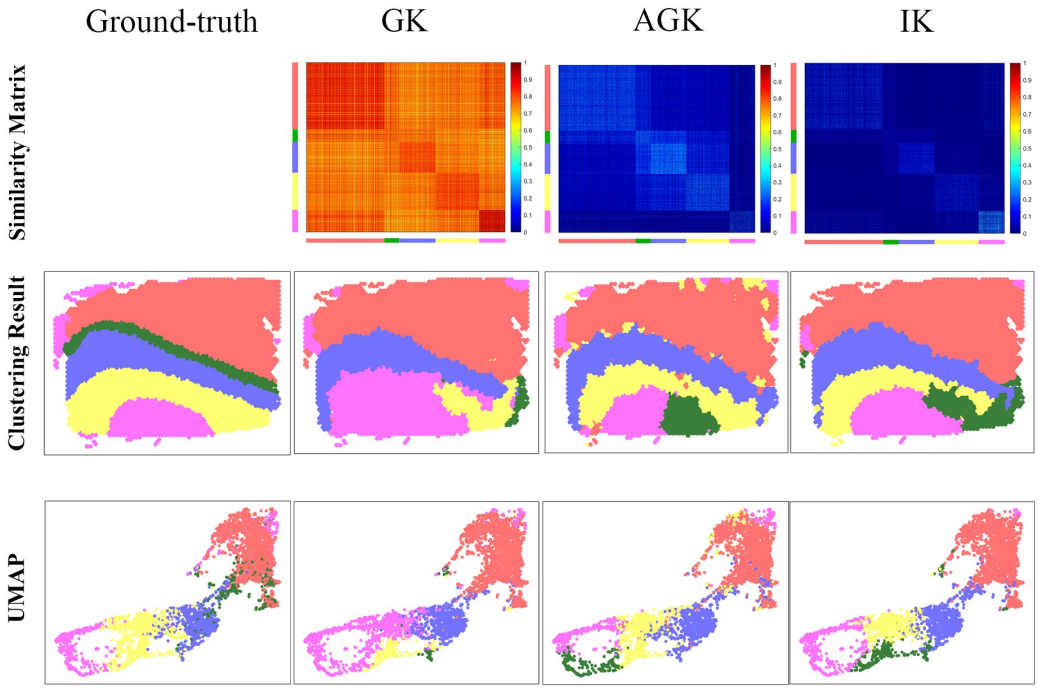

Figure S7. Results of the further ablation studies on three kernels and their associated KBC clustering outcomes in the original space and the reduced 2-D space by UMAP derived from the graph embedded space on the simulated SVGs (tissue slice 151672 of DLPFC dataset). GKBC, AGKBC and IKBC have ARI = 0.6, 0.57 and 0.68, respectively; and NMI = 0.56, 0.58 and 0.63, respectively.

The second example shown in Figure S7 has similar outcomes as in the first example (shown in Figure S6). One key exception is: AGKBC produced a slightly different clustering outcome: the pink cluster was split into two while the yellow cluster was correctly clustered, and it has encroached into the largest red cluster in quite a few scatter small regions.

Supplemental Text S4

KBC clustering of an example single cell resolution dataset

Our clustering method KBC and all other stand-alone clustering methods (like Kmeans & Mclust) are applicable to datasets generated by spatial transcriptomics technologies with cellular resolutions. Here we validated KBC's ability in identifying tissue structures on the mouse olfactory bulb (Chen et al. 2022), a widely used model tissue with the laminar organization. This SRT dataset was generated by Stereo-seq, a newly emerging spatial transcriptomics technology that could achieve the cellular or subcellular spatial resolution by DNA nanoball patterned array chips (Chen et al. 2022). The Stereo-seq data of mouse olfactory bulb tissue was downloaded from: https://github.com/JinmiaoChenLab/SEDR_analyses. Xu et al. 2024 has annotated the laminar organization of coronal mouse olfactory bulb in the DAPI-stained image, containing the rostral migratory stream (RMS), granule cell layer (GCL), internal plexiform layer (IPL), mitral cell layer (MCL), external plexiform layer (EPL) and olfactory nerve layer (ONL) (Figure S8A). Following the same preprocessing procedure used for other datasets, we normalized the raw data with SCTransform (Choudhary et al. 2022), select a set of spatially variable genes (SVGs) with SPARK-X (J Zhu et al. 2021). We then conducted the experiments on the processed data.

As shown in Figure S8, the results from KBC mirrored the laminar organization well and the identified regions matched the annotated layers. As shown in Figure S8C, KBC clearly recognized the narrow tissue structure MCL, compared to Stagate, which was further validated by the expression of mitral cell marker Gabra1. Note that, for this dataset, KBC and SpatialPCA produced similar results.

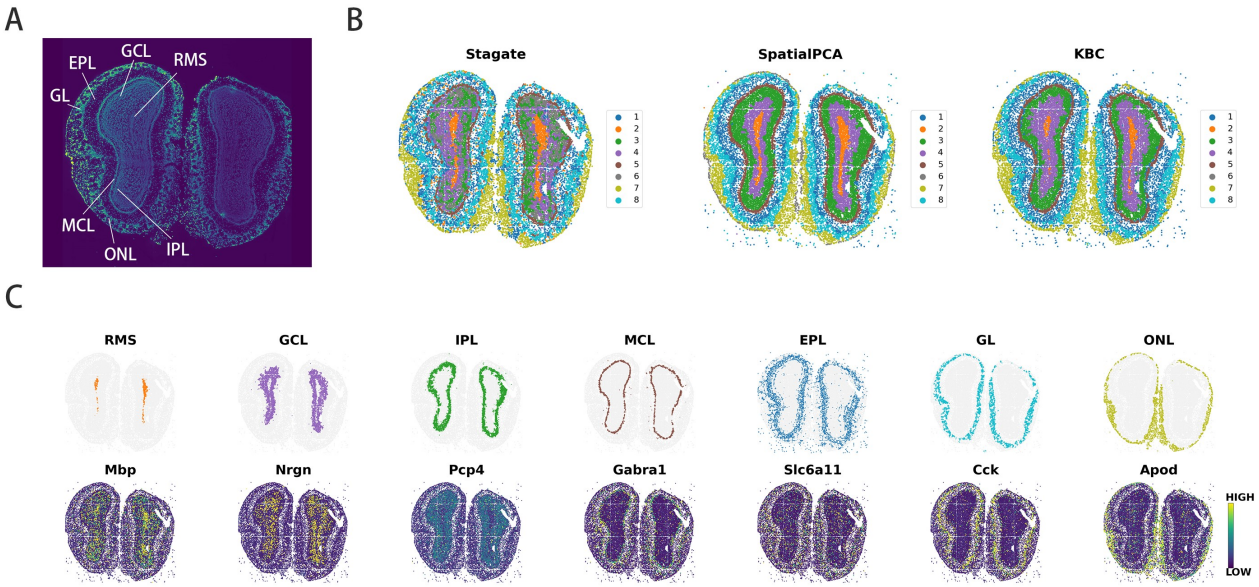

Figure S8. Application of KBC to the mouse olfactory bulb Stereo-seq data. **A**, Laminar organization of the mouse olfactory bulb annotated using the DAPI-stained image. **B**, Spatial domains identified by SpatialPCA and KBC in the mouse olfactory bulb Stereo-seq data. KBC uses the WL embedding here. **C**, Visualization of the spatial domains identified by KBC (first row) and the corresponding marker gene expressions (second row).

References

Bandaragoda TR, Ting KM, Albrecht D, Liu FT, Zhu Y, and Wells JR. 2018. Isolation-based anomaly detection using nearest neighbour ensembles. *Comput Intell*. **34**: 968–998.

Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Qiu X, Yang J, Xu J, Hao S, et al. 2022. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*. **185**: 1777–1792.

Choudhary S and Satija R. 2022. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**: 27.

Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M, et al. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci*. **24**: 425–436.

McInnes L, Healy J, and Melville J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Shang L and Zhou X. 2022. Spatially aware dimension reduction for spatial transcriptomics. *Nat Commun*. **13**: 7203.

Ting KM, Zhu Y, and Zhou ZH 2018. Isolation kernel and its effect on SVM. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Xu H, Fu H, Long Y, Ang KS, Sethi R, Chong K, Li M, Uddamvathanak R, Lee HK, Ling J, et al. 2024. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*. **16**: 12.

Zelnik-Manor L and Perona P 2005. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*.

Zhu J, Sun S, and Zhou X. 2021. SPARK-X: Non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol.* **22**: 1–25.

Zhu Y and Ting KM. 2021. Improving the Effectiveness and Efficiency of Stochastic Neighbour Embedding with Isolation Kernel. *J Art Intell Res*. **71**: 667–695.