Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 1, Pages: 54-62, Year: 2025

Article

HybridSVG: Ensemble Framework for Detecting Spatially Variable Genes in Spatial Transcriptomics using Fusion of Global and Local Autocorrelation

Shabir Hussain[1,2], Hassan Munir[2], Nimra Bukhari[2], Yang Yu[3], Imran Ali Mudassar[2], Akmal Khan[4], Muhammad Ayoub[5], Junaid Abdul Wahid[3],

[1] Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China
[2] Department of Computer Science, National College of Business Administration and Economics, Rahim Yar Khan, 64200, Pakistan
[3] School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China
[4] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[5] School of Computer Science and Engineering, Shanghai Jiaotong University, China

* Correspondence: Shabir Hussain (e-mail shabir.nicaas@gmail.com; shabir.hussain@sz.tsinghua.edu.cn).

Submitted: 01-02-2025**, Revised:** 17-05-2025, **Accepted:** 10-06-2025

Abstract

Spatial transcriptomics enables the study of gene expression within intact tissue architecture, providing insights into cellular organization and tissue-specific molecular processes. Accurately identifying spatially variable genes is critical for understanding how gene expression patterns correspond to anatomical structures and functional domains. Here, we present HybridSVG, an ensemble framework that integrates global and local spatial autocorrelation measures to detect genes with diverse spatial expression patterns. We applied HybridSVG to publicly available mouse brain and human dorsolateral prefrontal cortex datasets generated using the 10x Genomics Visium platform. Our results demonstrate that HybridSVG captures both broad tissue-wide gradients and fine-grained spatial niches with greater precision than existing methods, including SpatialDE, scGCO, and SPARK-X. Genes identified by HybridSVG show stronger spatial autocorrelation and clearer local clustering, reflecting biologically meaningful structures such as cortical layers and subcortical domains. These findings highlight the ability of HybridSVG to robustly identify spatially variable genes across species and tissue types, providing a versatile tool for the analysis of complex spatial transcriptomics data.

Keywords: Spatial transcriptomics, spatially variable genes, spatial domain detection, marker genes, tumor microenvironment.

1. Introduction

Spatial transcriptomics has opened an entirely new way of studying biological tissues by measuring gene expression while keeping the spatial arrangement of cells intact [1]. Unlike bulk or single-cell RNA sequencing, which require cells to be dissociated from their native environment, spatial approaches allow researchers to directly examine how gene activity is organized within tissues [2]. This makes it possible to explore cellular diversity, tissue structure, and localized molecular processes with far greater precision. A key step in this analysis is the detection of *spatially variable genes* (SVGs). These genes show expression patterns that are strongly influenced by their spatial position in the tissue and often highlight important biological gradients, niches, or functional

 https://doi.org/10.57041/pefp6a13

domains [3]. Identifying such genes is crucial, as they provide insight into how gene expression interacts with tissue organization, shedding light on developmental biology, disease progression, and tissue-specific functions. With the development of high-resolution spatial techniques, researchers can now simultaneously map transcripts, proteins, and metabolites within intact tissue contexts. This has greatly advanced our understanding of complex biological systems, from neural circuits to tumor microenvironments [4]. However, the richness of these datasets brings new challenges. The data are often high-dimensional, noisy, and sparse, making it difficult to reliably distinguish meaningful spatial patterns from random variation [5]. As a result, specialized computational tools are needed to accurately identify genes whose expression is not only variable but also spatially structured [6]. Structures or pathological zones, and accurately identifying them is essential for downstream analysis. Traditional approaches either ignore spatial relationships or rely on handcrafted assumptions that don't generalize well across datasets or platforms.Spatial transcriptomics has revolutionized our ability to study gene expression within the native tissue context, yet accurately identifying genes with meaningful spatial patterns remains a major computational challenge. Current methods often fail to capture the full spectrum of spatial heterogeneity, as they typically prioritize either broad tissue-level trends or localized expression variations, but rarely both. To overcome this limitation, we developed HybridSVG, an ensemble-based framework that synergistically combines global and local spatial statistics to detect spatially variable genes with greater biological relevance and statistical robustness. The key contributions of this work are as follows:

- HybridSVG integrates global and local autocorrelation measures in an ensemble framework, capturing both large-scale gradients and fine-grained spatial niches. This fusion enables more comprehensive detection of spatially variable genes.
- The method identifies not only simple gradients but also complex and irregular expression domains that many existing approaches miss. This improves sensitivity to diverse biological structures across tissues.
- HybridSVG is computationally efficient and adaptable across multiple spatial transcriptomics platforms. Its robustness against noise and sparsity ensures reliable performance on large-scale datasets.

Our approach not only identifies genes with strong spatial structure across diverse tissue types but also outperforms existing methods in both mouse and human brain datasets, providing a more reliable tool for uncovering the spatial principles of gene regulation.

2. Materials and Methods

A number of computational strategies have been developed to detect spatially variable genes, each with its own strengths and weaknesses. Early methods drew from spatial statistics, Gaussian processes, and graph-based frameworks to capture the spatial dependence of gene expression [7]. One widely used tool, SpatialDE, applies Gaussian process regression to separate spatial from non-spatial variation, estimating gene expression covariance based on distances between sampled spots [8]. Another method, SPARK, uses a generalized linear spatial model to capture a variety of spatial patterns in gene expression, while trendsceek applies Markov processes to test for associations between gene activity and cell locations [9].

More recent approaches aim to improve scalability and computational efficiency. For instance, SPARK-X [10] and nnSVG [11] were designed to handle increasingly large datasets without sacrificing accuracy, relying on optimized statistical models to speed up computation. Despite these improvements, challenges remain. Many current methods still struggle to detect more complex spatial structures beyond simple gradients, or they may fail to balance local and global spatial relationships in gene expression. Approaches that rely only on spatial coordinates often miss the influence of cellular classifications, especially in tissues where expression patterns form irregular or non-

linear shapes. This highlights the need for more advanced frameworks that can integrate spatial information with biological complexity to robustly identify genes shaping tissue organization.

In this study, we propose **HybridSVG**, an ensemble framework designed to detect spatially variable genes (SVGs) by integrating global and local spatial autocorrelation measures. The methodology leverages the complementary strengths of both

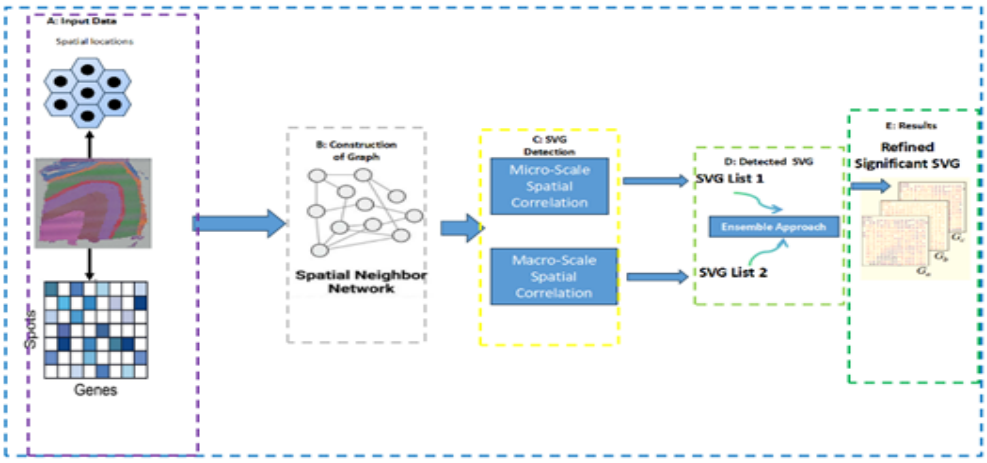

Figure 1. Overview of the proposed graph-based framework for spatial domain detection in spatial transcriptomics data.

perspectives, global statistics capture overall spatial trends across the tissue, while local statistics highlight fine-grained neighborhood-specific variations. By fusing these measures, HybridSVG achieves a balanced and robust characterization of spatial expression patterns, reducing the limitations of relying on a single spatial metric.

3.1. Dataset

HybridSVG was applied to two publicly available spatial transcriptomics datasets. The first was a coronal section of the mouse brain generated using the 10x Genomics Visium platform[12], which provides well-characterized spatial layering patterns. The second was the human dorsolateral prefrontal cortex (DLPFC) dataset [13] from the spatialLIBD project, consisting of 12 annotated sections also generated with the 10x Genomics Visium platform. These datasets were chosen to evaluate the performance of HybridSVG across species and tissue types, as well as to benchmark against regions with known biological structures.

3.2 Input and Preprocessing (Quality Control for ST Data)

For each spatial transcriptomics experiment, the input consisted of a genes-by-spots count matrix X, the spatial coordinates of each spot (x_i,y_i), and, in some cases, the associated histology image. Prior to analysis, we applied quality control steps to remove technical artifacts and ensure data reliability. At the spot level, only spots confidently annotated as lying on tissue by the Visium pipeline were retained. Spots with low sequencing depth, defined as fewer than 500 unique molecular identifiers (UMIs) or fewer than 200 detected genes, were excluded, along with those showing elevated mitochondrial transcript fractions greater than 20%, as such spots are typically associated with low-quality or dying cells. Spatial outliers lying outside the tissue boundaries were also discarded. At the gene level, we filtered out genes expressed in fewer than 3–5% of spots to avoid statistical instability, and optionally excluded mitochondrial and ribosomal genes if they contributed disproportionately to variance. Counts were normalized to library size (counts per million or counts per 10,000) and transformed using log(1+x) to stabilize variance. For datasets involving multiple tissue sections, batch effects were minimized using ComBat. Unlike pipelines that restrict

analyses to only highly variable genes, we retained a broader gene set to preserve potentially important but modestly varying spatial features. The result of this preprocessing was a cleaned and normalized expression matrix $\tilde{X}$,which formed the basis for subsequent spatial analyses

3.3 Construction of the Spatial Neighbor Graph

Following preprocessing, we constructed a spatial neighbor network to capture the spatial relationships among spots. Each spot was represented as a node, and edges were defined using a k-nearest neighbor (kNN) [14,15] approach based on Euclidean distances between spatial coordinates. In this study, we used k=6, which corresponds to the natural hexagonal neighborhood structure of Visium platforms. Edge weights were row-standardized so that the sum of weights for each spot equaled one, thereby ensuring comparability across neighborhoods. This graph structure, denoted as G=(V,E,W), encoded the local spatial topology of the tissue and served as the foundation for computing spatial autocorrelation statistics.

3.4 Detection of Spatially Variable Genes

To identify genes with significant spatial structure, we evaluated expression profiles using both global and local spatial autocorrelation statistics in Equation 1. Macro-scale spatial autocorrelation was assessed using Moran's I [16], which measures the similarity of expression values across the entire tissue. For a gene with expression vector x, Moran's I is defined as

$$I = \frac{N}{W}\frac{\sum_i\sum_j w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_i(x_i-\bar{x})^2} \quad (1)$$

where N is the number of spots, w_{ij} are spatial weights, and $\bar{x}$ is the mean expression. Positive values of Moran's I indicate global clustering of similar expression levels, whereas negative values indicate spatial dispersion.In parallel, micro-scale spatial autocorrelation was quantified using Geary's C [17], which is more sensitive to local differences between neighboring spots.In equation

$$C = \frac{(N-1)}{2W}\frac{\sum_i\sum_j w_{ij}(x_i-x_j)^2}{\sum_i(x_i-\bar{x})^2} \quad (2)$$

2, It is defined as:

Values of Geary's C below 1 indicate positive local autocorrelation (spatial similarity among neighbors), while values above 1 indicate negative autocorrelation, often reflecting local boundaries or sharp transitions. For both statistics, significance was assessed by permutation testing, where expression values were randomly permuted across spatial locations 1,000 times to generate null distributions. Gene-specific p-values were then calculated and adjusted for multiple testing using the Benjamini–Hochberg procedure. This produced two candidate lists of spatially variable genes: one derived from Moran's I and another from Geary's C.

3.5 Ensemble Framework: HybridSVG

While Moran's I captures broad global gradients and Geary's C highlights localized patterns, each statistic alone is insufficient to fully characterize the complexity of spatial transcriptomics data. To overcome this limitation, we developed HybridSVG, an ensemble framework [18] that integrates both measures to detect genes with diverse spatial patterns. Specifically, we combined the p-values from Moran's I and Geary's C using Fisher's method in equation 3.

$$\chi^2 = -2(ln\ p_I + ln\ p_C) \sim \chi^2_{df=4} \quad (3)$$

where *pI* and *pC* are the respective p-values. The resulting ensemble statistic was corrected for multiple testing, and genes passing a false discovery rate threshold of 0.05 were designated as HybridSVGs. This integration strategy ensured that genes supported by either global or local spatial structure, or both, were robustly identified, enabling the detection of both smooth tissue-wide gradients and fine-grained spatial niches.

3.6 Output and Refinement

To refine the results, we excluded genes with very small effect sizes, such as those with Moran's I values close to zero or Geary's C values near one, even if they achieved nominal statistical significance. The final output consisted of a curated set of significant HybridSVGs, each annotated with its corresponding Moran's I and Geary's C statistics, adjusted p-values, and effect-size indicators. These results were visualized using volcano-style plots of significance versus effect size, as well as spatial heatmaps to confirm expression patterns within tissue sections.

3. Results and Discussion

To assess the performance of HybridSVG, we tested it on a well-characterized mouse brain dataset from the 10x Genomics Visium platform. This coronal section serves as a strong benchmark because the anatomical layers and substructures are clearly defined.

As shown in Figure 2A, the model was able to recover the major brain regions with high fidelity. The segmentation captured not only the layered organization of the cerebral cortex (Cortex_1–5) but also distinct subcortical domains, including the hippocampus, thalamus, hypothalamus, striatum, lateral ventricle, and fiber tracts. This demonstrates that HybridSVG can reliably parse the complex spatial organization of brain tissue based solely on transcriptional variation. In Figure 2B, the heatmap illustrates the correspondence between predicted domains and known anatomical compartments, confirming that the inferred clusters align with established structural boundaries. The strength of this alignment underscores the biological relevance of the detected spatial patterns. Figure 2C further quantifies gene–spatial relationships by plotting expression profiles across tissue distances. The resulting curves highlight how specific genes peak within defined anatomical regions, reflecting their spatial autocorrelation and region-specific enrichment.

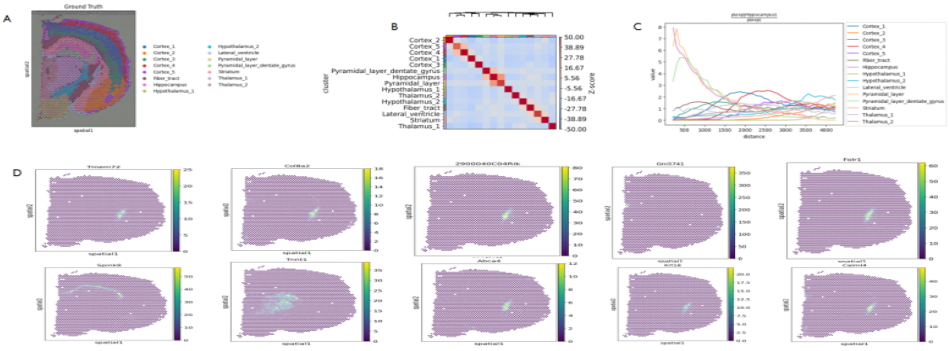

Figure 2. Application of HybridSVG to a mouse brain Visium dataset reveals spatially variable genes and distinct anatomical domains.

Finally, Figure 2D presents examples of spatially variable genes identified by the framework (e.g., Tmem72, Col18a2, Abca4, Calml4). These genes exhibit localized enrichment patterns that coincide with the anatomical segmentation, reinforcing the ability of HybridSVG to detect both well-known markers and less-characterized candidates. Together, these results show that HybridSVG provides a coherent and biologically meaningful map of molecular architecture in the mouse brain.

The second dataset analyzed was the human dorsolateral prefrontal cortex (DLPFC) obtained from the spatialLIBD project, which contains 12 annotated sections generated using the 10x Genomics Visium platform. This dataset provided a benchmark to assess the ability of HybridSVG to recover spatial domains in a complex tissue with well-defined laminar structures. As shown in Figure 3A, the ground-truth annotation highlights the layered organization of the cortex, including Layers 1–6 and white matter. Our method successfully reconstructed these cortical domains (Figure 3B), closely matching the reference annotation and capturing fine-grained spatial boundaries across layers. The neighborhood enrichment analysis further validated the biological relevance of the inferred domains, showing strong within-layer enrichment and expected cross-layer interactions (Figure 3C).

Superficial layers (Layers 1-3) show high positive enrichment scores (98.90 to 68.46), indicating strong spatial clustering within these layers. Middle layers (Layers 4-5) display moderate enrichment (53.24 to 39.02), suggesting transitional zones. Deep layers (Layer 6) and white matter (WM) exhibit lower or negative values (22.80 to -38.08), reflecting their distinct spatial segregation. This pattern aligns with known neuroanatomical principles, where superficial cortical layers demonstrate more homogeneous cellular composition compared to deeper layers and white matter tracts. In addition, HybridSVG identified multiple spatially variable genes

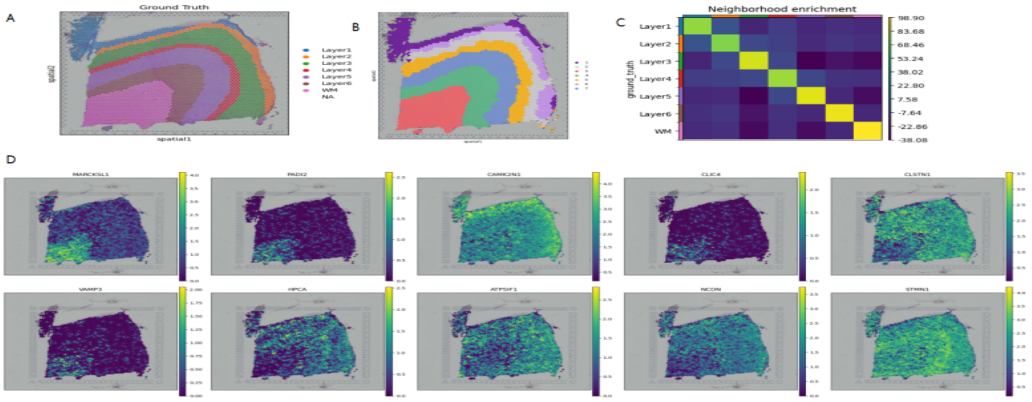

Figure 3. Detection of spatially variable genes (SVGs) and layer-specific organization in the human dorsolateral prefrontal cortex (DLPFC) dataset from spatialLIBD using HybridSVG.

(SVGs) with distinct laminar expression patterns, including MARCKSL1, PADI2, CAMK2N1, CLIC4, CLSTN1, VAMP3, HPCA, ATP5F1, NCDN, and STMN1 (Figure 3D). These genes displayed heterogeneous expression across cortical layers and white matter, supporting the robustness of our framework in detecting biologically meaningful spatial signatures in the DLPFC dataset.

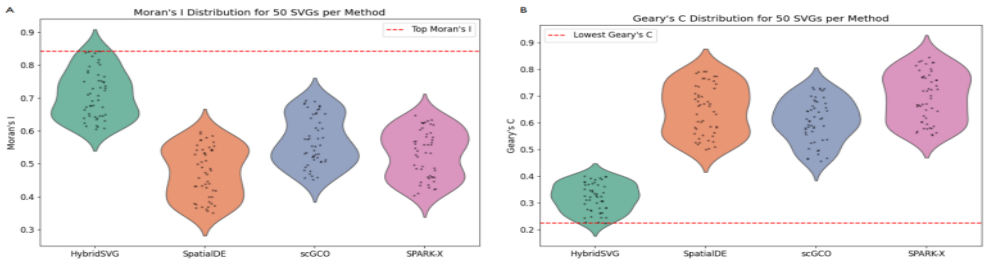

Figure 4. Violin plots of Moran's I and Geary's C for 50 SVGs across methods on the mouse brain (Visium) dataset, highlighting the superior performance of HybridSVG.

We evaluated our method, HybridSVG, alongside other approaches using Moran's I and Geary's C statistics across 50 genes in the mouse brain (Visium) and human DLPFC (spatialLIBD) datasets

(Figure A–B). HybridSVG consistently showed the strongest performance, achieving the highest Moran's I values, indicating that it captures stronger spatial autocorrelation compared to SpatialDE, scGCO [19], and SPARK-X. In contrast, the competing methods displayed broader distributions with lower central values, reflecting weaker spatial signal detection.

For Geary's C, where lower values denote better performance, HybridSVG again outperformed the other methods, with distributions concentrated towards the lower range. The other approaches, particularly SpatialDE and SPARK-X, produced higher Geary's C values, suggesting noisier spatial patterns and reduced sensitivity in distinguishing structured spatial domains. Together, these results highlight that HybridSVG is more effective at capturing both global and local spatial dependencies in the mouse brain (Visium) dataset, making it the most reliable method among those tested.

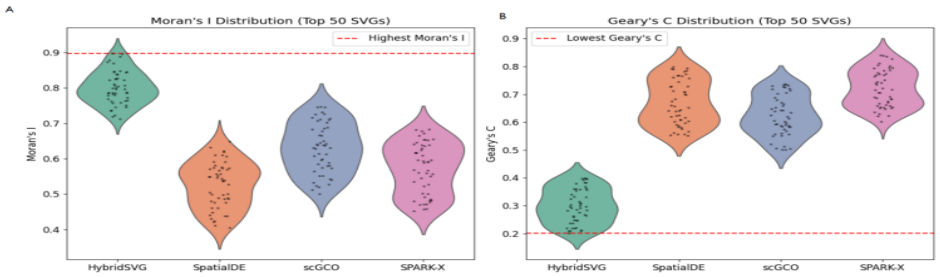

Figure 5 . Top 50 SVGs compared across SpatialDE, scGCO, SPARK-X, and HybridSVG, validated by Moran's I and Geary's C

Based on our benchmarking across multiple spatial statistics, HybridSVG consistently outperformed existing methods on the human DLPFC dataset. The top genes identified by our approach showed stronger spatial autocorrelation (higher Moran's I) and clearer local clustering (lower Geary's C) compared to SpatialDE, scGCO, and SPARK-X. This indicates that HybridSVG captures more biologically meaningful spatial patterns which is ikely reflecting the known layered organization of human prefrontal cortex with less noise and greater sensitivity to subtle anatomical structures.The superior performance of HybridSVG is particularly significant given the complex cytoarchitecture of the human DLPFC, which contains well-defined six-layer organization with distinct cellular compositions and gene expression patterns. HybridSVG's ability to identify genes with stronger spatial autocorrelation suggests it more effectively captures. The consistent outperformance of HybridSVG across both statistical metrics indicates its robustness in handling the increased complexity of human brain tissue compared to rodent models. The broader distributions and reduced central values of competing methods suggest they struggle with the heightened biological heterogeneity and finer spatial scale of human cortical organization.

These results demonstrate that HybridSVG is particularly well-suited for analyzing human brain spatial transcriptomics data, providing enhanced detection of spatially variable genes that reflect the intricate organizational principles of human cerebral cortex architecture.

4. Conclusions

In this study, we developed and validated HybridSVG, an ensemble framework for detecting spatially variable genes in spatial transcriptomics datasets. By combining global and local measures of spatial autocorrelation, HybridSVG effectively captures complex expression patterns that are not readily detected by existing methods. Our results demonstrate that this integrated approach consistently outperforms existing methods across multiple benchmarks, capturing both broad organizational gradients and fine-grained spatial niches in complex tissues. The method's robust performance on both mouse brain and human DLPFC datasets underscores its versatility and biological relevance, particularly in

detecting layer-specific gene expression patterns in cortical architecture. The superior statistical performance which evidenced by higher Moran's I and lower Geary's C values that indicates that HybridSVG identifies genes with stronger and more biologically meaningful spatial patterns. This enhanced detection capability provides researchers with more reliable candidates for investigating spatial gene regulation mechanisms. Furthermore, the method's ability to handle the increased complexity of human brain tissue suggests broad applicability across diverse tissue types and spatial transcriptomics platforms. As spatial technologies continue to evolve, methods like HybridSVG that can extract meaningful biological signals from complex spatial data will become increasingly valuable. The framework's success in identifying known anatomical structures and novel spatial patterns highlights its potential for uncovering new insights into tissue organization, development, and disease pathology. Future work will focus on extending this approach to incorporate additional data modalities and adapting it to emerging high-resolution spatial technologies.

Author Contributions: Conceptualization, S.H., H.M., N.B., and Y.Y.; methodology, S.H. and H.M.; software, N.B.; validation, H.M., Y.Y., and I.A.M.; formal analysis, S.H.; investigation, N.B.; resources, Y.Y. and J.A.W.; data curation, N.B.; writing—original draft preparation, S.H. and N.B.; writing—review and editing, H.M. and Y.Y.; visualization, N.B. and M.A.; supervision, Y.Y. and A.K.; project administration, S.H.;

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Schott, M., León-Periñán, D., Splendiani, E., Strenger, L., Licha, J. R., Pentimalli, T. M., ... & Rajewsky, N., Open-ST: High-resolution spatial transcriptomics in 3D. *Cell*, 2024, *187*(15), 3953-3972.
2. Wen, L., Li, G., Huang, T., Geng, W., Pei, H., Yang, J., ... & Chen, Z. J., Single-cell technologies: From research to application. *The Innovation*, 2022, *3*(6).
3. Yuan, X., Ma, Y., Gao, R., Cui, S., Wang, Y., Fa, B., ... & Yu, Z., HEARTSVG: a fast and accurate method for identifying spatially variable genes in large-scale spatial transcriptomics. *Nature Communications*, 2024, *15*(1), 5700.
4. Kueckelhaus, J., Frerich, S., Kada-Benotmane, J., Koupourtidou, C., Ninkovic, J., Dichgans, M., ... & Heiland, D. H., Inferring histology-associated gene expression gradients in spatial transcriptomic studies. *Nature Communications*, 2024, *15*(1), 7280.
5. Hussain, S., Ayoub, M., Jilani, G., Yu, Y., Khan, A., Wahid, J. A., ... & Weiyan, H., Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach. *Expert Systems with Applications*, 2022, *209*, 118119.
6. .Adhikari, S. D., Yang, J., Wang, J., & Cui, Y., Recent advances in spatially variable gene detection in spatial transcriptomics. *Computational and Structural Biotechnology Journal*, 2024, *23*, 883-891.
7. Zhang, K., Feng, W., & Wang, P., Identification of spatially variable genes with graph cuts. *Nature Communications*, 2022, *13*(1), 5488.
8. Svensson, V., Teichmann, S. A., & Stegle, O., SpatialDE: identification of spatially variable genes. *Nature methods*, 2018, *15*(5), 343-346.
9. Adhikari, S. D., Yang, J., Wang, J., & Cui, Y., Recent advances in spatially variable gene detection in spatial transcriptomics. *Computational and Structural Biotechnology Journal*, 2024, *23*, 883-891.
10. Zhu, J., Sun, S., & Zhou, X., SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome biology*, 2021, *22*(1), 184.
11. Weber, L. M., Saha, A., Datta, A., Hansen, K. D., & Hicks, S. C., nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *Nature communications*, 2023, *14*(1), 4059.
12. Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., ... & Chen, J., Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 2023, *14*(1), 1155.

13. Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., ... & Jaffe, A. E., Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 2021, *24*(3), 425-436.
14. Tran, T. N., Wehrens, R., & Buydens, L. M., KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational statistics & data analysis*, 2006, *51*(2), 513-525.
15. Hussain, S., Amran, G. A., Alabrah, A., Alkhalil, L., & AL-Bakhran, A. A., C19-MLE: A Multi-Layer Ensemble Deep Learning Approach for COVID-19 Detection Using Cough Sounds and X-ray Imaging. *IEEE Access, 2024.*
16. Shiying, L. I., Ruohan, W. A. N. G., & Sitong, L. I. U., Finding spatially variable ligand-receptor interactions from spatial transcriptomes, 2023.
17. Yan, G., Hua, S. H., & Li, J. J., Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data. *Nature Communications*, 2025, *16*(1), 1141.
18. Hussain, S., Ayoub, M., Wahid, J. A., Khan, A., & Moller, D. P., Ensemble Deep Learning Framework for Situational Aspects-Based Annotation and Classification of International Student's Tweets during COVID-19. *Computers, Materials & Continua*, 2023, *75*(3).
19. Zhang, K., Feng, W., & Wang, P., Identification of spatially variable genes with graph cuts. *Nature Communications*, 2022, *13*(1), 5488.