





Please cite the Published Version

Wang, Tianqi, Zhu, Huitong , Zhou, Yunlan, Ding, Weihong, Ding, Weichao , Han, Liangxiu  and Zhang, Xueqin  (2024) Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. *Communications Biology*, 7 (1). 1351 ISSN 2399-3642

DOI: <https://doi.org/10.1038/s42003-024-07037-0>

Publisher: Springer

Version: Supplemental Material

Downloaded from: <https://e-space.mmu.ac.uk/636364/>

Usage rights:  [Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Additional Information: The version of record of this article, first published in *Communications Biology*, is available online at Publisher's website: <http://dx.doi.org/10.1038/s42003-024-07037-0>

Data Access Statement: All datasets analyzed in this paper can be downloaded in raw form from the original publication. Specifically, the first dataset is human dorsolateral prefrontal cortex data³⁶ captured using 10 Visium technology and can be downloaded from <http://research.libd.org/spatialLIBD/>. The second dataset is the mouse embryo E9.5 spatial transcriptomic data³⁷ obtained with Stereo-seq technology and can be downloaded from <https://db.cngb.org/stomics/mosta/>. The third dataset is human breast cancer data obtained from the public 10 Genomics database and can be downloaded from <https://www.10xgenomics.com/resources/datasets/humanbreast-cancer-block-a-section-1-1-standard-1-1-0>. The fourth dataset is mouse brain tissue data obtained from the public 10 Genomics database. The mouse brain tissue contains anterior and posterior regions. Here, we use mouse brain anterior data, which can be downloaded from <https://www.10xgenomics.com/resources/datasets>. The last dataset is de novo NEPC and ARPC data³¹ obtained from the public 10 Genomics database, which can be downloaded from <https://db.cngb.org/stomics/datasets/STDS0000227>. The data used in this study have been uploaded to Zenodo and is freely available at: <https://zenodo.org/records/13731512>. All source data underlying the graphs and charts are presented in Supplementary Data. An open-source Python implementation of the GAAEST toolkit is accessible at <https://github.com/tqwang743/GAAEST-main>.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Supplementary Information

GAAEST's supplementary experiments on mouse embryo E9.5 dataset

According to the original annotation of mouse embryo E9.5 dataset, we first selected 12 as the number of clusters to test the domain recognition ability of different methods. The results, as depicted in Supplementary Figure 1, demonstrated that GAAEST exhibited the most superior performance, with its ARI value about 10% higher than the second-best method. Upon careful observation, it became evident that the domain contours identified by stLearn lacked clarity and exhibited poor continuity. The Heart region was not fully recognized by STAGATE, GraphST, SEDR, and SpaceFlow. SpaGCN failed to identify the connective tissue region. In contrast, GAAEST demonstrated superior domain recognition results, aligning more closely with the provided annotations and exhibiting clear contour segmentation. Furthermore, in order to achieve a more detailed tissue segmentation with higher resolution, we continuously increased the number of clusters from 12 to 24. Through experimentation, we determined that the optimal clustering results were obtained when using 20 clusters, as shown in Supplementary Figure 2.

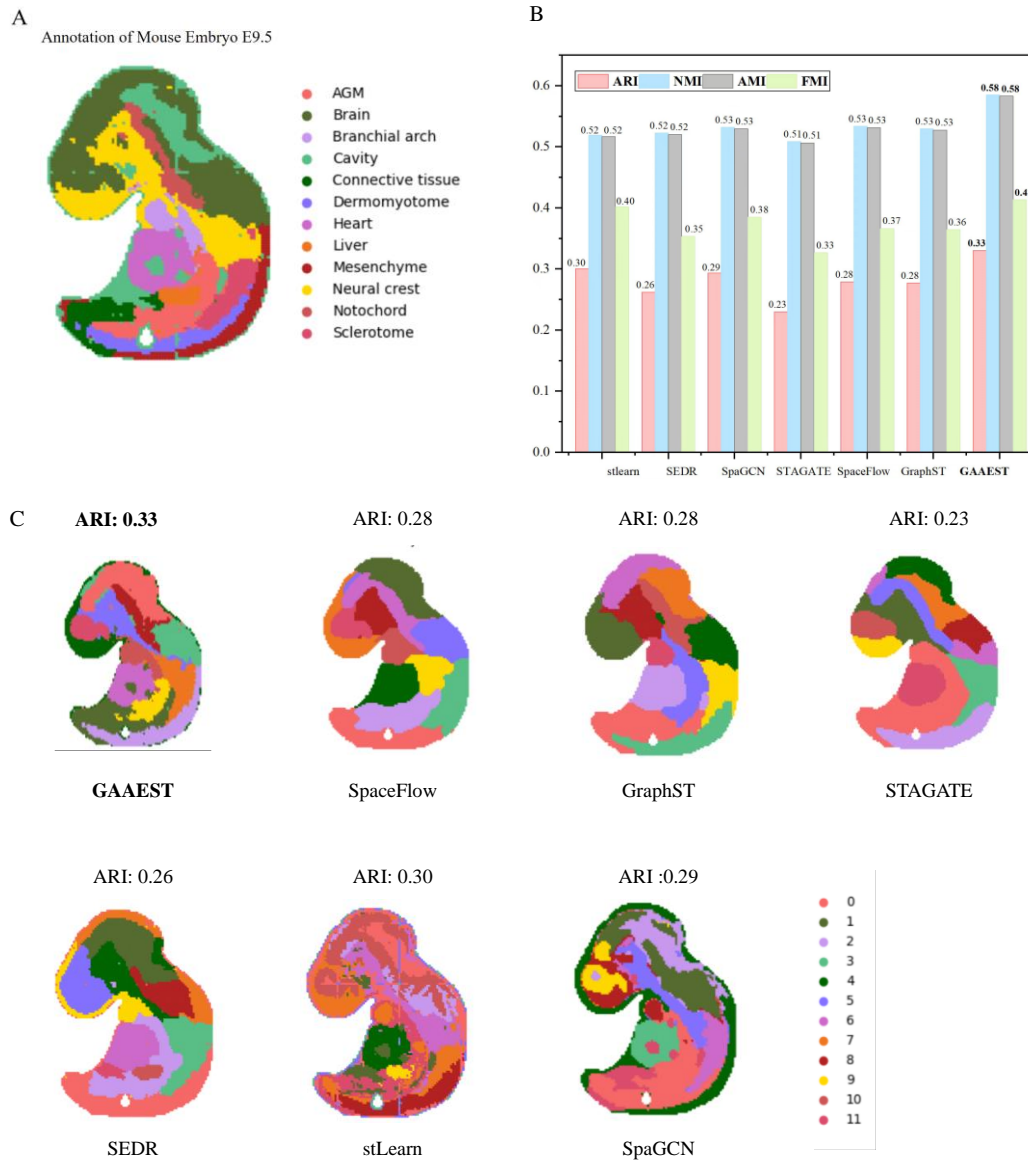
GAAEST's supplementary experiments to evaluate training time

We conducted an experiment on the human breast cancer dataset to demonstrate the operation time of every method. As with the other experiment parameters in human breast cancer dataset, we set the cluster numbers to 10 and 20, respectively. The experimental results are shown in Supplementary Table 4. The experimental results revealed notable differences in the training times among the considered methods when varying the number of clusters. Specifically, when the number of clusters was set to 10, SpaGCN exhibited the shortest training time, while stLearn required the longest duration. Comparatively, our proposed method achieved a substantial reduction in training time, being 71% shorter than stLearn. On the other hand, when the number of clusters was increased to 20, GraphST demonstrated the shortest training time, whereas stLearn exhibited the longest duration. Remarkably, our method outperformed stLearn by achieving a remarkable 74% reduction in training time. The observed reductions in training time underscore the potential of our approach in accelerating the analysis of clustering tasks, thus contributing to enhanced computational efficiency and overall productivity.

GAAEST's supplementary experiments to validate parameter design

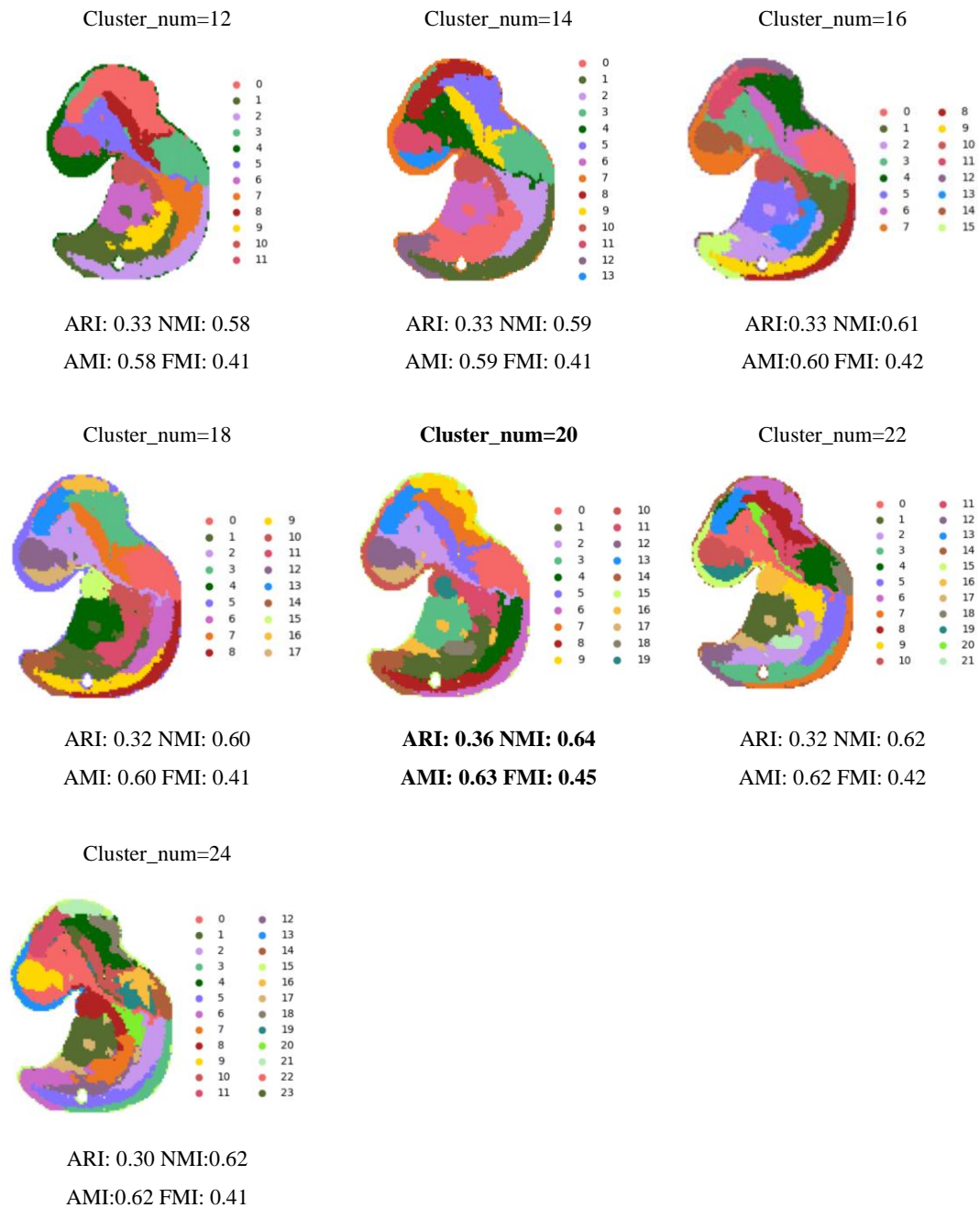
The parameters in Eq. (16) and (17) are of the most importance, so we conduct the relevant experiments to show how to determine the combination of these parameters in GAAEST. The specific experimental results are shown in Supplementary Table 5 and 6, it can be seen that the best performance was achieved when the $\beta:\alpha$ was set to 10:1 and $\lambda_1:\lambda_2:\lambda_3$ was set to 1:1:1.

Supplementary Figure 1. Comparison of spatial domains identified by GAAEST and six comparative methods on E9.5 mouse embryo data when cluster number was set to 12.



(A) Tissue domain annotation of mouse embryo E9.5 data obtained from the original Stereo-seq study. **(B)** The value of ARI, NMI, AMI, and FMI for GAAEST and comparative methods. **(C)** Spatial clustering visualization of GAAEST and six comparative methods.

Supplementary Figure 2. Comparison of clustering performance by GAAEST on E9.5 mouse embryo data as the number of clusters continued to increase from 12 to 24.



Supplementary Table 1. Summary of related work

Reference	Autoencoder structure		Self-supervised contrastive learning			Gene expression reconstruction	The embedding used for clustering	Clustering Methods
	Encoder	Decoder	Local	Global	Context			
			location-based	feature-based	feature-based			
SEDR (2021)	GCN, FCN	Inner Product, FCN	×	×	×	✓	EF	DEC
SpaGCN (2021)	GCN	—	×	×	×	×	EF	DEC
SpaceFlow (2022)	GCN	—	×	✓	×	×	EF	Leiden
STAGATE (2022)	GAT	GAT	×	×	×	✓	RF	Mclust
GraphST (2023)	GCN	GCN	×	✓	×	✓	RF	Mclust
RESEPT (2022)	GCN	Inner Product	×	×	×	×	—	—
Ours	GAT	FCN	✓	✓	✓	✓	RF	Mclust

The embedding used for clustering column: EF represents extracted feature by encoder, RF is reconstructed gene expression by decoder.

Clustering Methods: DEC is deep embedded clustering.

Supplementary Table 2. Results of GAAEST ablation experiments on the human breast cancer dataset with different clustering numbers

	Clustering number = 10				Clustering number = 20			
	ARI ↑	NMI ↑	AMI ↑	FMI ↑	ARI ↑	NMI ↑	AMI ↑	FMI ↑
GAAEST	0.677	0.714	0.711	0.717	0.611	0.704	0.698	0.640
w/o CFCL	0.649	0.707	0.705	0.695	0.506	0.698	0.693	0.543
w/o LLCL	0.645	0.703	0.700	0.687	0.507	0.650	0.644	0.544
w/o GFCL	0.582	0.688	0.686	0.635	0.523	0.700	0.694	0.559
w/o feature reconstruction	0.236	0.445	0.440	0.364	0.281	0.453	0.443	0.337

Supplementary Table 3. Results of using different autoencoder on the human breast cancer dataset with different clustering numbers

Encoder	Decoder	Clustering number = 10				Clustering number = 20			
		ARI↑	NMI↑	AMI↑	FMI↑	ARI↑	NMI↑	AMI↑	FMI↑
GAT	FCN	0.677	0.714	0.711	0.717	0.611	0.704	0.698	0.640
GCN	GCN	0.652	0.708	0.705	0.695	0.559	0.680	0.674	0.593
GAT	GAT	0.649	0.693	0.690	0.694	0.569	0.661	0.655	0.603

Supplementary Table 4. Training time (seconds) of related work on human breast cancer datasets.

	Cluster_num = 10	Cluster_num = 20
SpaGCN	95.83	111.74
stLearn	428.94	448.02
SEDR	259.37	178.17
SpaceFlow	139.32	157.91
GraphST	104.68	104.46
STAGATE	101.92	105.33
GAAEST	123.55	117.74

Supplementary Table 5. Related experiments on human breast cancer datasets to verify the selection of parameters $\beta:\alpha$

$\beta:\alpha$	Cluster_num = 10				Cluster_num = 20			
	ARI \uparrow	NMI \uparrow	AMI \uparrow	FMI \uparrow	ARI \uparrow	NMI \uparrow	AMI \uparrow	FMI \uparrow
1:1	0.551	0.645	0.641	0.602	0.547	0.656	0.650	0.580
5:1	0.614	0.679	0.676	0.664	0.533	0.678	0.673	0.568
10:1	0.676	0.713	0.711	0.716	0.611	0.703	0.698	0.640
15:1	0.647	0.707	0.705	0.690	0.604	0.698	0.693	0.634
20:1	0.644	0.701	0.699	0.689	0.574	0.691	0.686	0.606
1:5	0.571	0.637	0.634	0.626	0.484	0.623	0.616	0.522
1:10	0.433	0.562	0.557	0.510	0.420	0.569	0.561	0.464
1:15	0.481	0.593	0.589	0.545	0.409	0.565	0.557	0.455

Supplementary Table 6. Related experiments on human breast cancer datasets to verify the selection of parameters $\lambda_1:\lambda_2:\lambda_3$

$\lambda_1:\lambda_2:\lambda_3$	Cluster_num = 10				Cluster_num = 20			
	ARI \uparrow	NMI \uparrow	AMI \uparrow	FMI \uparrow	ARI \uparrow	NMI \uparrow	AMI \uparrow	FMI \uparrow
1:1:1	0.676	0.713	0.711	0.717	0.611	0.704	0.698	0.640
2:1:1	0.597	0.677	0.674	0.651	0.606	0.681	0.676	0.647
1:2:1	0.592	0.675	0.672	0.648	0.538	0.670	0.664	0.572
1:1:2	0.530	0.671	0.668	0.601	0.585	0.676	0.671	0.616
1:2:2	0.587	0.677	0.674	0.644	0.607	0.686	0.681	0.638
2:2:1	0.590	0.674	0.671	0.645	0.512	0.662	0.656	0.556
1:2:1	0.548	0.683	0.680	0.618	0.559	0.681	0.676	0.594