Contrastive learning approaches for spatial transcriptomics clustering

Simone Colli

Corresponding author(s). E-mail(s): simone.colli@studenti.unipr.it;

1 Introduction

The spatial organization of cells within a tissue is related to its biological function, and perturbations in this architecture are characteristic of many pathological conditions. [1] For many years, transcriptomic analyses have offered profound insights into cellular molecular machinery; however, conventional methods not includes tissue information, resulting in the loss of critical spatial context. The emergence of spatial transcriptomics (ST) has revolutionized this field by merging high-throughput sequencing with precise spatial coordinates. [2]

A primary objective in the analysis of ST data is the accurate identification of spatial domains. These are defined as distinct regions within a tissue composed of spots or cells that exhibit both spatial proximity and analogous gene expression profiles. The precise delineation of such domains is of paramount importance for comprehending tissue structure, developmental biology, and disease pathogenesis. [3]

Initial computational strategies involved the adaptation of non-spatial clustering algorithms, including K-means and Louvain [4, 5], which operate exclusively on gene expression data.

Although these methods provided some utility, they frequently yielded domains that were fragmented and lacked biological coherence, a consequence of their failure to incorporate the essential spatial component of the data.

A substantial advancement was achieved with the application of Graph Neural Networks (GNNs). Through the representation of tissue as a graph—wherein spots constitute the nodes and spatial proximity defines the edges—methods such as SpaGCN [6], STAGATE [7] and SEDR [8] employ graph-based architectures to learn low-dimensional embeddings that encapsulate both molecular and spatial features. These embeddings, in turn, offer a more robust foundation for clustering and the identification of coherent tissue structures. More recent innovations have been driven by self-supervised contrastive learning, an advanced deep learning technique that refines learned embeddings by training the model to discriminate between analogous and non-analogous spots. [9] This is accomplished by optimizing the model to minimize the distance between the representations of neighboring spots while maximizing the distance between those of non-proximal or dissimilar spots within the embedding space. This methodology, as implemented in leading models like GraphST [10], GAAEST [11], and stCluster [12], has been demonstrated to yield more informative and discriminative representations.

2 State-of-the-Art

The current situation of spatial transcriptomics clustering methods combining different techniques and different data modalities, such as gene expression, spatial coordinates, and histology images, to identify a spatially coherent and biologically meaningful set of clusters. In recent years, the application of artificial intelligence and machine learning techniques has emerged as a powerful tool. [12] Prominent example of such methods include GAAEST [11], stCluster [12] and GraphST (only spatial informed clustering module) [10].

2.1 GAAEST

GAAEST proposes as a generalised deep learning method that integrates both spatial location details and gene expression data from transcriptomics. [11]

This framework consists of six main components: (1) data preprocessing, (2) neighbour graph construction, (3) data argumentation, (4) auto-encoder, (5) self-supervised contrastive learning for embedding refinement, and (6) spatial clustering.

As first the data preprocessing, the raw gene expression matrix is log-transformed, standardised to library size, and scaled to have unit variance and zero mean. Then only the 3000 highly variable genes (HVGs) are selected for the following steps.

Next, the neighbour graph construction is performed, where the spatial location information about spots are converted into a neighbour graph using the k-nearest neighbor algorithm [13]. Here, nodes represent the spots and, undirected and unweighted, edges connect a spot to its k-closest neighbours based on Euclidean distance. The resulting adjacency matrix is then regularised to ensure that all spots have a more equal influence, independent of their degree.

After that, the data argumentation is applied to create a permuted version of the graph. Here, the gene expression vectors are randomly permuted without changing the original graph structure. This permuted graph is then used to train the model in a contrastive learning task.

Next, the auto-encoder is applied to embed the gene expression data with the spatial location information into a low-dimensional feature embedding. The autoencoder component consists of two main parts: the encoder and the decoder. The encoder consists of a two-layer graph attention network (GAT) and takes as input the adjacency matrix of the neighbour graph and the gene expression, to produce the low-dimensional embedding. The decoder consists of two fully connected layers, whose primary goal is to reconstruct the original gene expression matrix from the low-dimensional embedding.

The model is trained to minimise the reconstruction loss (L_{RECON}), calculated as the mean square error between the original and the reconstructed gene expression vectors. To refine the learned representations, a self-supervised contrastive learning strategy on three levels is applied. The first level is the local location-based contrastive learning (LLCL), which is designed to enhance the network's focus on the unique spatial properties of each spot. It works by distinguishing a specific spot's embedding from the embeddings of other spots in different locations, using the original and permuted graphs to define positive (same spot on different graphs) and negative pairs (different spots on different graphs). The second level is the global feature-based contrastive learning (GFCL), which strengthens the model against the local noise. This strategy is used to learn global structural properties by improving the mutual information between the feature representation of a single spot and a global summary of the entire graph. The third level is the context feature-based contrastive learning (CFCL), here this strategy, leveraging on the literature [14], relay on the behaviour that same tissue type tends to exhibit similarities in terms of marker genes and morphological structures. This module learn the context feature by maximising the mutual information between the individual spot representation and the cluster-level summary. The total contrastive learning loss (L_{CL}) is a weighted sum of the losses obtained from the three-level contrastive learning strategy. The final model loss (L_{TOTAL}) is a combination of the reconstruction loss and the contrastive learning loss.

Finally, the spatial clustering is performed on the learned representations using the Mclust algorithm [15], which assumes the data originates from a mixed Gaussian distribution, to assign spots to different spatial domains.

To further refine the clustering results, the assignment for each spot is corrected, if more than 50% of a spot's neighbors belong to a single cluster, the spot is reassigned to that majority cluster.

2.2 stCluster

stCluster proposes as novel method that integrates graph contrastive learning and multi-task learning to refine informative representation for spatial transcriptomic data. [12]

This framework is composed of three main steps. (1) As first spatial gene expression profile is encoded using a GAT-based graph encoder, which captures the spatial dependencies and interactions among spots within the tissue. (2) Next, a combined model optimisation strategy that utilizes graph contrastive learning and multi-task learning. The contrastive learning enhances the discrimination capability of the generated representations. The multi-task learning enables the model to obtain simultaneously Multiple related tasks and improve the overall performance of the algorithm. (3) Finally, a clustering algorithm is applied to a clustering An algorithm to detect spatial domains based on the learned representations.

In addition to these steps, stCluster also incorporates more phases. At the beginning, a preprocessing step is performed. Here, spots located outside of the main tissue area are removed, and remaining spots are standardized in order to maintain the number of genes per spot up to 3000. At the end of the preprocessing, all raw gene expression profiles are then log-transformed and scale-normalized.

After the preprocessing step, the spots are represented as a spatially adjacent graph (SAG), where nodes represent the spots and undirected and unweighted edges represent the spatial relationships between the spots. Two spots are considered neighbours if they are within a certain Euclidean distance threshold.

Next, the graph is constructed, and the encoder, composed of two layers, one of graph attention networks (GAT), and one of the fully connected layers, is applied to encode the SAG and the gene expression profiles into a low-dimensional representation. The first layer of GAT is used to capture the spatial dependencies and interactions among spots, while the second layer of fully connected layer is used to further refine the representation. During the training, the model parameters are optimised by applying a combined optimization strategy that utilises graph contrastive learning at each iteration and periodically fine-tuning multi-task learning.

The optimization process done by graph contrastive learning uses two altered version, referred to as "views", of the original SAG, by randomly pruning edges. Then each view is encoded by the encoder, and the model is trained to maximise the similarity between the representation of the same spot in the two views, while minimising the similarity between the representations of different spots. The multi-task learning is used to refine the learned representations by simultaneously optimising the model for three tasks. First is the adjacency matrix reconstruction (AMR) task, which aims to reconstruct the adjacency matrix of the SAG from the learned representations. Second is the gene expression reconstruction (GER) task, which aims to Reconstruct the gene expression profiles from the learned representations. Third is the spatial domain prediction (SDP) task, which applies a DEC algorithm [16] to refine the clustering performance based on the learned representations.

In the final step, the optimised latent representations learned during the training are used to identify the spatial domains. The framework feeds these representations into the Mclust clustering algorithm [15], which models the data using a Gaussian finite mixture model. Here, the algorithm requires the number of clusters to be specified in advance, and for datasets where the number of clusters is unknown, the authors suggest using the Louvain community detection algorithm as an alternative.

2.3 GraphST

GraphST proposes as graph self-supervised contrastive learning method that fully exploits spatial transcriptomics data, by combining graph neural networks with self-supervised contrastive learning.

GraphST is composed of three modules, only the first one is relevant for this review. This module is composed of four main steps: (1) data pre-processing and argumentation, (2) GNN encoder for latent representation learning, (3) self-supervised contrastive learning for representation refinement, and (4) decoder for gene expression reconstruction.

As first raw gene expression counts are first log-transformed and normalized by library size. Then, the top 3000 highly variable genes (HVGs) are selected to be the input of the model. Next, the spatial location information are used to construct a neighbour graph using a predefined number of neighbors. The spots are represented as nodes and the undirected and unweighted edges represent the spatial relationships between the spots. For a given spot, its neighbors are defined as the k-nearest spots based on the Euclidean distance computed from spatial location information. For subsequent contrastive learning task, is generated a corrupted neighbour graph by randomly shuffling the gene expression vectors among the spots without changing the original graph structure.

Then a GNN encoder is used to learn spot representations that capture the informative parts of the gene expression profile and spatial locations. Specifically, the encodere is a graph convolutional network (GCN) that generates a latent representation for each spot by iteratively aggregating features from its neighbours. This latent representation is then passed to a symmetric decoder, which aims to reconstruct the original gene expression matrix.

To make the the learned representations more informative and discriminative, the model is optimized using a self-supervised contrastive learning (SCL) strategy. This strategy is designed to explicitly capture the local spatial context of each spot. The local spatial context of a spot consists in the aggregation of neighbors representations, which is considered as the spot's neighborhood microenvironment [17]. The positive pair are defined as the representation of a spot and its neighborhood microenvironment, represented as a vector of averanged embedding of its immediate neighbors. The negative pairs are defined as the representations of the same spot and the neighborhood of that spot in the microenvironment from the corrupted neighbour graph. An important part of this approach is to maximize the mutual information of positive pairs while minimizing the mutual information of negative pairs. To distinguish between positive and negative pairs, the model employs a binary cross-entropy (BCE) loss. For improved stability, this loss is applied symmetrically to both the original and the corrupted graphs The final objective function is then computed as a weighted sum of these contrastive losses and the self-reconstruction loss.

After training, the mclust algorithm is applied to the reconstructed gene expression matrix to identify the final spatial domains. The framework also includes an optional refinement step, where a spot's cluster assignment can be corrected based on the majority label of its neighbours.

3 Conclusion

The accurate identification of spatial domains is a fundamental task in the analysis of spatial transcriptomics data, essential for unveiling tissue architecture and underlying biological mechanisms. This review has examined the evolution of computational approaches for spatial clustering, highlight more recent technique like self-supervised contrastive learning.

The analyzed methods, including GAAEST, stCluster, and GraphST, represent the state-of-the-art and demonstrate the power of integrating spatial information and gene expression through deep learning architectures. A common thread among these approaches is the use of graph-based autoencoders to learn low-dimensional latent representations of tissue spots. The key innovation lies in the use of contrastive learning strategies to refine these representations, making them more informative and discriminative. While they share this fundamental principle, the methods differ in their specific optimization strategies: GAAEST employs a three-level contrastive learning approach (local, global, and contextual); stCluster combines contrastive learning with a multi-task learning framework for simultaneous refinement across multiple biological objectives; GraphST focuses on capturing the local context by defining positive and negative pairs based on a neighborhood ”microenvironment.”

These advanced approaches have shown significant improvements in clustering accuracy across a wide range of datasets and technological platforms, enabling a more precise and biologically coherent delineation of tissue structures, from the brain to developing embryos. Their ability to produce robust representations translates into better identification of domains with sharp boundaries and greater correspondence with manual annotations and marker gene patterns.

References

[1] Rao, A., Barkley, D., França, G.S., Yanai, I.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)

[2] Hwang, B., Lee, J.H., Bang, D.: Single-cell rna sequencing technologies and bioinformatics pipelines. Experimental & molecular medicine **50**(8), 1–14 (2018)

[3] Moses, L., Pachter, L.: Museum of spatial transcriptomics. Nature methods **19**(5), 534–546 (2022)

[4] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature biotechnology **33**(5), 495–502 (2015)

[5] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

[6] Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee, E.B., Shinohara, R.T., Li, M.: SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature methods **18**(11), 1342–1351 (2021) https://doi.org/10.1038/s41592-021-01255-8

[7] Dong, K., Zhang, S.: Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nature communications **13**(1), 1739 (2022)

[8] Xu, H., Fu, H., Long, Y., Ang, K.S., Sethi, R., Chong, K., Li, M., Uddamvathanak, R., Lee, H.K., Ling, J., *et al.*: Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Medicine **16**(1), 12 (2024)

[9] Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L.: Graph contrastive learning

with adaptive augmentation. In: Proceedings of the Web Conference 2021, pp. 2069–2080 (2021)

[10] Long, Y., Ang, K.S., Li, M., Chong, K.L.K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A., *et al.*: Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nature Communications **14**(1), 1155 (2023) https://doi.org/10.1038/s41467-023-36796-3

[11] Wang, T., Zhu, H., Zhou, Y., Ding, W., Ding, W., Han, L., Zhang, X.: Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. Communications Biology **7**(1), 1351 (2024) https://doi.org/10.1038/s42003-024-07037-0

[12] Wang, T., Shu, H., Hu, J., Wang, Y., Chen, J., Peng, J., Shang, X.: Accurately deciphering spatial domains for spatially resolved transcriptomics with stcluster. Briefings in Bioinformatics **25**(4), 329 (2024) https://doi.org/10.1093/bib/bbae329

[13] Zhang, Z.: Introduction to machine learning: k-nearest neighbors. Annals of translational medicine **4**(11), 218 (2016) https://doi.org/10.21037/atm.2016.03.37

[14] Zong, Y., Yu, T., Wang, X., Wang, Y., Hu, Z., Li, Y.: const: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. BioRxiv, 2022–01 (2022) https://doi.org/10.1101/2022.01.14.476408

[15] Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R journal **8**(1), 289 (2016) https://doi.org/10.32614/RJ-2016-021

[16] Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 478–487. PMLR, New York, New York, USA (2016). https://proceedings.mlr.press/v48/xieb16.html

[17] Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D.: Deep graph infomax. arXiv preprint arXiv:1809.10341 (2018) https://doi.org/10.48550/ARXIV.1809.10341