STdGCN: accurate cell-type deconvolution using graph convolutional networks in spatial transcriptomic data

Yawei Li[1,2], Yuan Luo[1,2,*]

[1] *Department of Preventive Medicine, Northwestern University, Feinberg School of Medicine, Chicago, IL 60611, USA*

[2] *Center for Collaborative AI in Healthcare, Northwestern University, Feinberg School of Medicine, Chicago, IL 60611, USA*

[*] Corresponding author. Email: yuan.luo@northwestern.edu (Yuan Luo)

Abstract

Spatially resolved transcriptomics performs high-throughput measurement of transcriptomes while preserving spatial information about the cellular organizations. However, many spatially resolved transcriptomic technologies can only distinguish spots consisting of a mixture of cells instead of working at single-cell resolution. Here, we present STdGCN, a graph neural network model designed for cell type deconvolution of spatial transcriptomic (ST) data that can leverage abundant single-cell RNA sequencing (scRNA-seq) data as reference. STdGCN is the first model incorporating the expression profiles from single cell data as well as the spatial localization information from the ST data for cell type deconvolution. Extensive benchmarking experiments on multiple ST datasets showed that STdGCN outperformed 13 published state-of-the-art models. Applied to a human breast cancer Visium dataset, STdGCN discerned spatial distributions between stroma, lymphocytes and cancer cells for tumor microenvironment dissection. In a human heart ST dataset, STdGCN detected the changes of potential endothelial-cardiomyocyte communications during tissue development. Our results demonstrate that STdGCN can serve as a robust and versatile tool for cell type deconvolution across multiple ST platforms and tissues. STdGCN is available as open source Python software at https://github.com/luoyuanlab/stdgcn.

Introduction

The development of spatially resolved transcriptomic technologies has improved our understanding of the spatial organization and heterogeneity of gene expression within a tissue [1-4]. Spatial transcriptomics (ST) technologies can process measurements of transcriptomes while maintaining the spatial structural information about the tissue, which has been applied in disease pathology research [5-7]. However, ST technologies still have limitations due to the trade-off between cell resolution and throughput. For example, many fluorescence in situ hybridization (FISH) based technologies cannot detect the full scope of the transcriptome within a single cell [8], which leaves computational challenges for imputing the

missing gene expression of FISH data. In addition, the sequencing spot/location of many sequencing-based methods do not have sufficient resolution at the single cell level, raising computational challenges for deconvolving cells detected within the same spatial spot/location. In contrast to ST technologies, scRNA-seq can detect many thousands of gene expressions within a single cell. Therefore, we can leverage the properties and advantages of scRNA-seq as a complementary tool to address some limitations of ST technologies. For example, restoring cell type decomposition within a spot/location is similar to using scRNA-seq to deconvolve bulk RNA-seq data. Indeed, a number of models have been developed for bulk RNA-seq data deconvolution [9-14], and these models could theoretically be used to model ST data. However, the cell number of a spot/location in ST data is far less than bulk RNA-seq data. The application of a bulk RNA-seq deconvolution method to such a small sample size usually results in noise from unrelated cell types, and this has been demonstrated by several benchmark analyses [15-17].

Recently, several models have been developed for deconvolving ST data using statistics-based or machine learning based algorithms [15-27]. For statistics-based algorithms, most studies assume that the gene transcript count follows a distribution, usually Poisson [18] or negative binomial [15, 17, 22, 26]. These methods first employed the scRNA-seq data to calculate the distribution coefficients of genes for each cell type, and then inferred cell type proportions using maximum-likelihood estimation or maximum a posteriori estimation. For machine learning based algorithms, studies usually integrate the predicted ST data with scRNA-seq data to predict cell type proportions directly through a neural network system, so that these models are not required an additional step to learn cell type gene signatures from scRNA-seq data [20, 21]. However, a common limitation of these models is that they independently deconvolved the spots in a given ST dataset, but ignored the neighborhood connections between spots. Indeed, studies have demonstrated that ST data can be isolated into several spatial domains [28, 29]. Integrating the spatial coordinate information between spots can calibrate the ST cell type decomposition learned from scRNA-seq and increase predictive accuracy.

In this study, we present the Spatial Transcriptomics deconvolution using Graph Convolutional Networks (STdGCN) for ST data cell-type deconvolution. Our model is based on graph convolutional networks (GCN), a widely used graph-based deep learning model. The highlight of STdGCN is that it is built on two link graphs -- an expression link graph representing the gene expression similarity between nodes and a spatial link graph representing the spatial connections between spots in the ST data. Incorporating expression and spatial graphs, STdGCN is able to accurately predict the cell type proportions of ST data. To test the performance of STdGCN, we benchmarked this model with 13 published models. STdGCN has shown its capability to outperform benchmarked models across different ST platforms.

Results

STdGCN description. STdGCN is designed for cell type deconvolution of ST data by using scRNA-seq data as the reference (**Fig. 1**). It hypothesizes that the ST data and scRNA-seq data have common cell types, and that the cell type gene transcript signatures for both ST data and scRNA-seq data are

similar. In short, the normalized gene expression of a spot can be computed as a mixture of cell types with different proportions. The first step of STdGCN is to identify the cell-type marker genes and randomly generate pseudo-spots using the scRNA-seq data (**Fig. 1A**). Then, it builds two link graphs to model the GCN pipeline (**Fig. 1B~1C**). The first link graph, i.e. the expression graph, is a hybrid graph consisting of three sub-graphs, a pseudo-spots internal graph, a real-spots internal graph, and a real-to-pseudo-spots graph. Each subgraph is constructed by mutual nearest neighbors (MNN) based on their expressed similarity between spots. The second link graph, i.e. the spatial graph, is built on the Euclidean distance between one real-spot and all other real-spots in ST data. When running STdGCN, the input feature matrix is forward propagated to the expression GCN layers and the spatial GCN layers, respectively. The outputs (Exp-feature and Spa-feature) of the two GCN are then concatenated into a single matrix by column, which is eventually forward propagated to the fully-connected layers to predict the cell type proportions in each spot. To train the model, we split the pseudo-spots into a training dataset and a validating dataset. Only pseudo-spots in the training dataset are back propagated and updated, whereas the validating dataset is used for the early-stopping purpose. In this way, the cell type proportions of real-spots can also be updated through the GCN pipeline with the learning of the pseudo-spots (**Fig. 1D**). The details are described in **Methods**.

Benchmarks of STdGCN in different datasets. To assess the performance of our model, we compared STdGCN to 13 existing published deconvolution models, including stereoscope, RCTD, SPOTlight, SpatialDWLS, Cell2location, DSTG, CellDART, DestVI, STdeconvolve, STRIDE, Tangram, BayesPrism and AdRoit. We selected four different single cell level with cell type annotations ST platforms (seqFISH [30], seqFISH+ [29], MERFISH [31] and slide-seq [32]) and used them to simulate multi-cellular synthetic ST datasets to test the performance of these models (**Fig. 2**, **Methods**). To compare the performance of the 14 models across the four datasets, we used both Jensen-Shannon divergence (JSD) and root-mean-square error (RMSE) to measure the distance between predicted proportions compared to the ground truth, and ranked these models based on their JSD and RMSE values, respectively. Considering all datasets have more than one slice, the performance of a model for a given dataset is calculated as the mean rank across all slices in the dataset.

Table 1 displays the comprehensive performance of the 14 models. The ranked results showed that STdGCN achieved the lowest average JSD and RMSE across all four ST platforms (tie for first place with RCTD in JSD of the MERFISH dataset). More specifically, for the seqFISH and seqFISH+ datasets, STdGCN was first ranked in both JSD and RMSE across all slices. In the meantime, STdGCN was ranked first among four out of the six slices under both metrics of the Slide-seq dataset. For the MERFISH dataset, STdGCN also achieved best performance among 66% slices in RMSE and 42% slices in JSD, respectively.

We next examined the performance of the models in spots with different numbers of cells. We split the spots into two groups -- a smaller group that the cell numbers within the spots were fewer or equal to five, and a larger group that the spots contained more than five cells. The average JSD and RMSE of the larger group was lower than in the smaller group across all datasets (P-value < 0.05, Wilcoxon test) for

most models (**Fig. 3A~3D**, **Fig. S1**), indicating that spots with smaller cell number are more difficult to be precisely deconvolved. Comparing the relative performance between two groups, the average ranks of most methods were consistent (**Table S1**). Even so, STdGCN still achieved the best JSD and RMSE scores in both the larger group and the smaller group, except for smaller group of the MERFISH dataset in the JSD metrics.

Cell-cell communication is an essential feature of multicellular organisms. Cell-cell communication is constrained by physical location in the cellular microenvironment and cellular interactions are likely to be spatially adjacent. Thus, accurate deconvolution facilitates the identification of co-localization and spatial interactions between cell types. To this end, we scrutinized the performance of the benchmarked models with different number of cell types in a single spot. Similarly, we split the spots into four groups: a singleton group (only one cell type within a spot), a doubleton group (two cell types within a spot), a tripleton group (three cell types within a spot) and a multiton group (at least four types of cells within a spot). Of the four groups, STdGCN still exhibited robust performance (**Fig. 3E~3H**, **Fig. S2**, **Table S2**), and in most circumstances, StdGCN achieved the lowest JSD and RMSE scores.

Lastly, we assessed the models' capability for predicting diverse cell types. To process the assessment of the predictive performance of a given cell type, we used a binary JSD to measure the distance between predicted proportions and the real proportions by summarizing the proportions of the remaining cell types for each spot (**Methods**). To measure the RMSE of the cell type, we only calculated the RMSE between predicted proportions and the real proportions across all spots of the given cell type in a slice. In terms of the average ranks of the 14 models, STdGCN also outperformed other models for most cell types across the four datasets (**Fig. 3I~3L**, **Fig. S3**). Note that based on the average ranks of all cell types, STdGCN was more specialized in predicting cell types with higher frequency (**Table S3~S4**).

Developing human breast cancer transcriptomic dataset. Having demonstrated the capability of STdGCN to identify cell-type proportions in synthetic multi-cellular ST datasets, we then investigated the performance of STdGCN on a real multi-cellular ST dataset from human breast cancer [33]. This study performed scRNA-seq (**Fig. 4A~4B**) and 10X Genomics Visium spatial transcriptomic sequencing of breast tumor samples. We used the scRNA-seq data as the reference to retrieve the cell type proportions for each 10X Genomics Visium spot. We compared the cell type proportions predicted by STdGCN with the authors' region annotation results. Most cell types are well mapped to expected locations (**Fig. S4~S9**).

In terms of the authors' annotation results, several regions were annotated as mixtures of cancer cells, stroma cells and lymphocytes. By using STdGCN, we were able to quantitatively dissect the tumor microenvironment (TME) for these tissues in high resolution. For example, though the overall proportions of cell types were similar between two patients CID4465 and CID4535 (**Fig. 4C**), the spatial structure indicated that their TMEs were excessively different. In the patient CID4535, lymphocytes were mainly located at the periphery of tumor cells, whereas they were highly mixed with cancer cells in patient CID4465, indicating a high rate of tumor-infiltrating lymphocytes (TILs) (**Fig. 4D~4G**). Studies have demonstrated that TIL density is a strong prognostic marker for survival in patients: when a locally developed breast cancer is treated with neoadjuvant chemotherapy, any occurrence of TILs is an obvious

predictor of the response [34-36]. Thus, STdGCN can be a useful tool for physicians evaluating the rates of TILs for a patient. Similarly, cell-cell interactions between stromal cells and cancer cells/lymphocytes have been reported to be important to the maintenance of TME [37]. On one hand, cancer cells that interact with stromal cells can create favorable physical or molecular signaling environments for tumor progression [38]. On the other hand, some stromal cells, including cancer-associated fibroblasts (CAFs) can recruit immune cells to the tumor tissues and influence their behavior towards cancer cells [39]. From the deconvolution results, we successfully identified extensive stroma-cancer and stroma-lymphocyte interactions in patients CID4290, 1160920F and 1142243F (**Fig. 4H, Fig. S4~S6**).

Developing human heart spatial transcriptomic dataset. To test the versatility of STdGCN, we then implemented STdGCN to investigate the spatial structure of a developing human heart spatial transcriptomics dataset [40]. This study used spatial transcriptomics [41, 42], scRNA-seq and *in situ* sequencing (ISS), to measure transcriptomic profiles in human embryonic heart tissues (**Fig. 5A~5D**). We used the scRNA-seq dataset as the reference to infer the cell-type composition in the spatial transcriptomics dataset. The spatial structures were well mapped between the predicted spatial transcriptomics dataset and the ISS dataset across most cell types (**Fig. 5A, Fig. S10**). For example, epicardial cells and epicardial derived cells were mainly on the outside of the heart tissue, and smooth muscle cells were in the aorta and pulmonary artery area. Besides, the distributions of cell-type marker genes were consistent with the predicted results (**Fig. 5E~5J, Methods**). Studies have reported that communication between endothelial cells and cardiomyocytes regulates are very important in cardiac development. Endothelial cells not only can transmit oxygenated blood supply to cardiomyocytes, but also provide local protective signals that promote cardiomyocyte organization and survival [43]. In order to examine the endothelial-cardiomyocyte communications in this dataset, we investigated the cell type co-localization score (**Methods**) of the three stages (PCW4.5-5, PCW6.5, PCW9) of heart tissue sections (**Fig. 5K**). The heatmap shows a higher co-localization score between endothelial cells (including "Endothelium / pericytes / adventitia" and "Capillary endothelium" in the dataset) and cardiomyocytes (including "Myoz2-enriched cardiomyocytes", "ventricular cardiomyocytes", "atrial cardiomyocytes" in the dataset) compared to the overall cell type level.

Of these cell types, the main difference between STdGCN expected results and ISS annotation results is the *Myoz2*-enriched cardiomyocytes and ventricular cardiomyocytes. In terms of the ISS dataset, *Myoz2*-enriched cardiomyocytes were observed in the ventricle and the atria areas with high proportion. However, our prediction results showed that most *Myoz2*-enriched cardiomyocytes were inferred as ventricular cardiomyocytes or atrial cardiomyocytes based on the location of *Myoz2*-enriched cardiomyocytes (**Fig. 5L~5O**). Note that besides STdGCN, previous studies showed that most models also mis-inferred *Myoz2*-enriched cardiomyocytes (Figure 3 in Chen et al. [44]). Indeed, the *Myoz2*-enriched cardiomyocytes are a subpopulation of cardiomyocytes in the healthy heart expressing *MYOZ2* [40, 45]. The UMAP plots of both the ISS dataset and the scRNA-seq dataset indicated that *Myoz2*-enriched cardiomyocytes have two sub-clusters, one is closer to ventricular cardiomyocytes and one is closer to atrial cardiomyocytes (**Fig. 5B and 5D**). In particular, we found that ventricular cardiomyocytes and

Myoz2-enriched cardiomyocytes were highly mixed in the UMAP plot of the ISS dataset, suggesting that the expression levels between the two cell types were highly consistent. Next, we investigated the cell type marker genes of the three types of cardiomyocytes (**Methods**) and found that 41.6% and 100% of *Myoz2*-enriched cardiomyocyte marker genes were also the atrial cardiomyocyte marker genes and the ventricular cardiomyocyte marker genes, respectively (**Fig. 5P**). In particular, *MYOZ2*, which was reported as the most significant marker gene of *Myoz2*-enriched cardiomyocytes (average log-fold change = 1.81, expressed in 99.0% ventricular cardiomyocytes, FDR ≈ 1.08×10^{-89}) [40, 45], was also the marker gene of ventricular cardiomyocytes (average log-fold change = 1.30, expressed in 92.7% ventricular cardiomyocytes, FDR ≈ 0) (**Methods**). In summary, the discrepancy may be due to the high similarity between *Myoz2*-enriched cardiomyocytes with other cardiomyocytes, but also indicates that current deconvolution models are not precise enough to distinguish between some subtypes.

Discussion

Spatial transcriptomic profiles have enabled a better understanding of tissue architecture at the cellular level. However, many current spatial transcriptional profiling technologies do not have sufficient single cell resolution. Thus, we propose STdGCN, a graph-based deep learning framework that leverages cell type profiles learned from single-cell RNA-seq to deconvolve the cell type mixtures in ST data. We benchmarked STdGCN on four ST datasets with 13 published ST deconvolution methods [15-27] and ranked the models based on their performance. Among the four datasets, STdGCN achieved the best performance in both JSD and RMSE across the four datasets (**Table 1**).

To improve the performance of cell type deconvolution, distinct designs were adapted based on the characteristics of ST datasets. First, in contrast with previous ST deconvolution models, STdGCN is the first model incorporating spatial structure information from ST data and expression profiles from scRNA-seq for cell type deconvolution. The advantage of this approach is that it leverages high similarity between spots within a spatial domain to calibrate the prediction based on the expression profile of scRNA-seq data. Next, to maximize the heterogeneity of the pseudo-spot pool, we developed a sampling-by-cell-type method to simulate the pseudo-spot. This method does not rely on the cell-type distributions of scRNA-seq by specifying equal weights for each cell type, which successfully addressed the scenario that rare cell types are unlikely to be randomly selected using traditional pseudo-spot simulation methods. In addition, we limited the maximum number of cell types within a single pseudo-spot, in order to mimic the characteristics of real ST data.

The modular design of STdGCN makes it flexible for users. Although we have provided a set of default parameters, to fit the heterogeneity of different ST platforms, STdGCN offers plenty of optional parameters for the user to refine the model based on their own datasets. For example, in the preprocessing module, STdGCN allows either a raw count matrix or a normalized expression matrix from ST data and scRNA-seq data as the input. It also enables the user to use their own genes of interest to process deconvolution. In the pseudo-spot simulation module, STdGCN allows the user to choose between a range of cell number and cell-type number within a spot based on the characteristics of the corresponding ST platform. The simulated spot number is also optional, which support the user to determine a balance

between better performance (i.e. more simulated spot number) and higher computational speed and with less memory. In the GCN module, the weights of the link graphs, the depth of both expression/spatial GCN layers and the multi-task fully-connected neural network layers, and the dimensionality of the hidden layers are all optional. Lastly, STdGCN offers both the csv table and the anndata format [46] as the output, benefitting the user to incorporate numerous downstream analysis tools, such as Scanpy, that are easy to use.

Despite the aforementioned merits, STdGCN still has limitations. First, the predictive performance of STdGCN varies across cell types. Specifically, the relative performance of STdGCN in cell types with lower proportions is not as good as in the cell types with higher proportions (**Table S3~S4**), suggesting that our current pseudo-spot generation method or the link graph construction model still needs to be improved. Thus, our future work will focus on ameliorating the effectiveness of the use of reference data. Second, as shown in **Fig. 5K~5N**, STdGCN cannot accurately distinguish cell types with very similar expression profiles. This may be a common issue for all deconvolution models. Boosting the marker gene selection might mitigate the issue and increase the model's performance.

In conclusion, STdGCN is the first proposed model that incorporates the spatial coordinates from spatial transcriptomics data and cell type profiles learned from single-cell RNA-seq data for deconvolving local cell type proportions in complex tissues. We expect that STdGCN could help to better dissect the spatial architecture of tissues and offer better support for cellular level downstream analysis.

Competing Interests

The authors declare that they have no conflict of interests.

References

1. Eng, C.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239 (2019).
2. Merritt, C.R. et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol* **38**, 586-599 (2020).
3. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
4. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018).
5. Berglund, E. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* **9**, 2419 (2018).
6. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38**, 333-342 (2020).
7. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer Res* **78**, 5970-5979 (2018).
8. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272 (2019).
9. Du, R., Carey, V. & Weiss, S.T. deconvSeq: deconvolution of cell mixture distribution in

sequencing data. *Bioinformatics* **35**, 5095-5102 (2019).
10. Tsoucas, D. et al. Accurate estimation of cell-type composition from gene expression data. *Nat Commun* **10**, 2975 (2019).
11. Wang, X., Park, J., Susztak, K., Zhang, N.R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).
12. Newman, A.M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782 (2019).
13. Gong, T. & Szustakowski, J.D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-1085 (2013).
14. Dong, M. et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform* **22**, 416-427 (2021).
15. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3**, 565 (2020).
16. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* **49**, e50 (2021).
17. Yang, T. et al. AdRoit is an accurate and robust method to infer complex transcriptome composition. *Commun Biol* **4**, 1218 (2021).
18. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517-526 (2022).
19. Dong, R. & Yuan, G.C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol* **22**, 145 (2021).
20. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform* **22** (2021).
21. Bae, S. et al. CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data. *Nucleic Acids Res* **50**, e57 (2022).
22. Lopez, R. et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol* (2022).
23. Sun, D., Liu, Z., Li, T., Wu, Q. & Wang, C. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res* **50**, e42 (2022).
24. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18**, 1352-1362 (2021).
25. Chu, T., Wang, Z., Pe'er, D. & Danko, C.G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* **3**, 505-517 (2022).
26. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* **40**, 661-671 (2022).
27. Miller, B.F., Huang, F., Atta, L., Sahoo, A. & Fan, J. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat Commun* **13**, 2339 (2022).
28. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375-1384 (2021).
29. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data.

316 *Nat Biotechnol* (2018).
317 30. Lohoff, T. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse
318 organogenesis. *Nat Biotechnol* **40**, 74-85 (2022).
319 31. Moffitt, J.R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic
320 preoptic region. *Science* **362** (2018).
321 32. Chen, H. et al. Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell Rep*
322 **37**, 109915 (2021).
323 33. Wu, S.Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**,
324 1334-1347 (2021).
325 34. Loi, S. et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient
326 Analysis of Early-Stage Triple-Negative Breast Cancers. *J Clin Oncol* **37**, 559-569 (2019).
327 35. Stanton, S.E. & Disis, M.L. Clinical significance of tumor-infiltrating lymphocytes in breast
328 cancer. *J Immunother Cancer* **4**, 59 (2016).
329 36. Zgura, A., Galesa, L., Bratila, E. & Anghel, R. Relationship between tumor infiltrating
330 lymphocytes and progression in breast cancer. *Maedica (Bucur)* **13**, 317-320 (2018).
331 37. Seager, R.J., Hajal, C., Spill, F., Kamm, R.D. & Zaman, M.H. Dynamic interplay between
332 tumour, stroma and immune system can drive or prevent tumour progression. *Converg Sci Phys*
333 *Oncol* **3** (2017).
334 38. Pietras, K. & Ostman, A. Hallmarks of cancer: interactions with the tumor stroma. *Exp Cell Res*
335 **316**, 1324-1331 (2010).
336 39. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16**, 582-598
337 (2016).
338 40. Asp, M. et al. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing
339 Human Heart. *Cell* **179**, 1647-1660 e1619 (2019).
340 41. Stahl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial
341 transcriptomics. *Science* **353**, 78-82 (2016).
342 42. Salmen, F. et al. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in
343 mammalian tissue sections. *Nat Protoc* **13**, 2501-2534 (2018).
344 43. Hsieh, P.C., Davis, M.E., Lisowski, L.K. & Lee, R.T. Endothelial-cardiomyocyte interactions
345 in cardiac development and repair. *Annu Rev Physiol* **68**, 51-66 (2006).
346 44. Chen, J. et al. A comprehensive comparison on cell-type composition inference for spatial
347 transcriptomics data. *Brief Bioinform* **23** (2022).
348 45. Gladka, M.M. et al. Single-Cell Sequencing of the Healthy and Diseased Heart Reveals
349 Cytoskeleton-Associated Protein 4 as a New Modulator of Fibroblasts Activation. *Circulation*
350 **138**, 166-180 (2018).
351 46. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data
352 analysis. *Genome Biol* **19**, 15 (2018).
353 47. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression
354 data. *Genome Biol* **22**, 78 (2021).
355 48. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-
356 sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-
357 427 (2018).
358 49. Kipf, T.N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks.
359 arXiv:1609.02907 (2016).

Methods

Benchmark data selection and preprocess. Herein, we selected four datasets from four single cell-level resolution ST platforms: seqFISH, MERFISH, seqFISH+ and slide-seq, to assess the performance of the benchmarked models. We resampled the datasets by splitting each slice into multiple square pixel areas. Cells within each square pixel area were merged into a synthetic spot. Since these ST platforms were single cell level, the exact cell type proportions for each synthetic spot were known. The single-cell level data was also used as the internal single cell reference.

The seqFISH+ dataset was obtained from Zhu et al. [29] (download by Giotto package [47]: getSpatialDataset(dataset = "seqfish_SS_cortex", method = "wget"), which detected mRNAs for 10,000 genes corresponding to a section of the mouse somatosensory (SS) region. We split the SS region into two slices based on the author's annotation, including a cortex slice and a sub-ventricular zone (SVZ) slice. For the cortex slice, we chose 400 × 400 square pixel area as a synthetic spot, whereas for the SVZ slice we chose 200 × 200 square pixel area. Any spot with cell number less than two was discarded. There were 109 cortex synthetic spots and 59 SVZ synthetic spots in the two slides after resampling, respectively.

The seqFISH dataset was obtained from Lohoff et al. [30] (https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/). This study detected mRNAs for 387 target genes in tissue sections of three mouse embryos (two slices for each embryo) at the 8–12 somite stage. In terms of the cell types annotated by the authors from single-cell transcriptome atlases, we removed the cell types with very small proportions (< 3%) for cell type deconvolution. The coordinates of the cells were stored in columns "x_global_affine" and "y_global_affine" in the file "metadata.Rds". For each slice, we selected the 0.1 × 0.1 square global_affine area as a synthetic spot. Any spot with cell number less than two was discarded. The final number of predictive synthetic spots for each slice after filtering ranged from 1279 to 2350.

The Slide-seq dataset was downloaded from Chen et al. [32] (https://www.dropbox.com/s/ygzpj0d0oh67br0/Testis_Slideseq_Data.zip?dl=0). This study captured the spatial gene expression patterns at near-single-cell resolution in three diabetes mouse testis slices and three wild type (WT) mouse testis slices. The coordinates of the cells were stored in columns "x" and "y" in the csv files starting with "BeadLocationsForR". For each slice, we selected the 80 × 80 square pixel area as a synthetic spot. Any spot with cell number less than two was discarded. The final number of predictive synthetic spots for each slice after filtering ranged from 2046 to 2615.

The MERFISH dataset was downloaded from Chen et al. [31] (https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248). This study detected mRNAs for a set of 155 target genes in the mouse preoptic region. Approximately 1.1 million cells were measured across 181 slices and 36 mice. Cell types with proportions smaller than 3% were removed for cell type deconvolution. The coordinates of the cells were stored in columns "xcoord" and "ycoord" in the file "Moffitt_and_Bambah-Mukku_et_al_merfish_all_cells.csv". For each slice, we selected the 50 × 50 square pixel area as a synthetic spot. Any spot with cell number less than two was discarded. The final number of predictive synthetic spots for each slice after filtering ranged from 938 to 1247.

Preprocess of the human breast cancer dataset. The human breast cancer 10X Genomics Visium dataset was obtained from Wu et al. [33]. The dataset includes six breast cancer patients, comprising two ER^+s and four TNBCs. When processing data deconvolution, we only considered the spots within the tissues. The coordinates of the spots were saved in the file "tissue_positions_list.csv". In this study, the authors also processed scRNA-seq from 26 breast tumor samples. To process the deconvolution, we used this scRNA-seq data as the reference.

Preprocess of the human heart dataset. The human heart spatial transcriptomics dataset was obtained from Asp et al. [40]. This dataset contains three sequencing platforms: spatial transcriptomics [41, 42], scRNA-seq, and *in situ* sequencing (ISS). The tissue sections were collected from human embryonic cardiac samples collected at three developmental stages in the first trimester: 4.5–5, 6.5, and 9 post-conception weeks (PCW). In this study, we employed STdGCN to infer the cell-type composition of the spatial transcriptomics data. This dataset consists of 19 slices (four slices in 4.5-5 PCW, nine slices in 6.5 PCW, and six slices in 9 PCW). When processing data deconvolution, we only considered the spots within the tissues. The coordinates of the spots can be obtained from their corresponding spot names. We used the scRNA-seq dataset provided in this study as the reference data for deconvolution. Since the spatial transcriptomics dataset does not have a single cell resolution, we cannot achieve the realistic cell type compositions for the spots. However, this study offers the differentially expressed gene (DEG) analysis results based on the scRNA-Seq data (Table S3 in [40]). To test the performance of STdGCN, we identified the cell-type marker genes from DEG analysis results and explored their distributions in the ST dataset. To filter cell-type marker genes, we only kept the genes with FDR (marked as "'p_val_adj'") < 0.01, average log-fold change (marked as "'avg_logFC'") > 1, fraction of cells expressing the genes in the cell type (marked as "pct.1") > 0.7, and fraction of cells expressing the genes in the remaining cell types (marked as "pct.2") < 0.3 from the downloaded DEG analysis table.

Benchmark metrics. To assess the cell type deconvolution performance for the algorithms across the synthetic real datasets, we selected the Jensen-Shannon divergence (JSD) and root mean square error (RMSE) as two metrics to measure the differences between the predicted cell type proportions and real proportions within each spot.

The JSD is a symmetric form of the Kullback-Leibler (KL) divergence. Denote P and Q are the proportional distributions of predicted and ground-truth results, respectively. The JSD can be expressed as:

$$\mathrm{JSD}(\mathrm{P}||\mathrm{Q}) = \frac{1}{2}D_{KL}(P||\frac{1}{2}(P+Q)) + \frac{1}{2}D_{KL}(Q||\frac{1}{2}(P+Q)) \qquad (1)$$

where D_{KL} is the Kullback-Leibler (KL) divergence, which can be expressed as:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \qquad (2)$$

Typically, the JSD is a nonnegative number, and a smaller JSD represents a higher similarity between the

predicted result and the ground-truth, indicating a better performance of the corresponding method.

To assess the performance of the models for cell type A, denote the predicted and ground-truth proportions of a spot are P(A) and Q(A). We considered the remaining cell types as a group defined as the non-A type with the proportions 1-P(A) and 1-Q(A), respectively. The equation (2) will be expressed as:

$$D_{KL}(P||Q) = P(A)\log\frac{P(A)}{Q(A)} + (1-P(A))\log\frac{\left(1-P(A)\right)}{\left(1-Q(A)\right)} \qquad (3)$$

Based on equation (3), we can eventually get the binary JSD between predicted proportions compared to the ground truth of the cell type A.

We used the following equation to calculate RMSE:

$$\mathrm{RMSE(P||Q)} = \sqrt{\frac{\sum_{i=1}^{K}(P(x_i)-Q(x_i))^2}{K}} \qquad (4)$$

where K is the number of cell types when assessing the performance of a model for deconvolving a spot, and where K is the number of spots within a slice when assessing the performance of the models for a given cell type. RMSE is a nonnegative number, and a smaller RMSE represents a lower dissimilarity between the predicted result and the ground truth.

Cell type co-localization score. Cell–cell interactions and communication can be processed via receptor–ligand pairs. Thus, identifying cell type co-localization provides potential cell-cell interactions within a tissue. For a ST slice, we develop a cell type co-localization score $\mathrm{S(A,B)}$ to assess the degree of cell type A and cell type B. $\mathrm{S(A,B)}$ can be calculated using the following functions:

$$\mathrm{S(A,B)} = \frac{\sum_{all\ spots} S_A = 1\ and\ S_B = 1}{\sum_{all\ spots} S_A = 1\ or\ S_B = 1}$$

where $S_A = 1$ denotes that the location has cell type A. The value of $\mathrm{S(A,B)}$ ranges between 0 and 1. With this index, higher $\mathrm{S(A,B)}$ indicates that the two cell types have a higher probability of co-localizing within a spot. Considering that the proportions of all cell types within a spot predicted by STdGCN are greater than 0, we used a threshold so that $S_A = 0$ when the predicted proportion of cell type A is less than the threshold. We used 0.15 as the threshold for the 10X Genomics Visium human breast cancer dataset co-localization analysis, and 0.1 for the developing human heart spatial transcriptomic dataset in terms of the characteristics of the two platforms.

Implementation of STdGCN

Single-cell data preprocessing and feature selection. The first step in processing STdGCN is to identify the cell type marker genes from scRNA-seq data. The input of the single-cell expression matrix is flexible and can be used for either the raw count matrix or the normalized matrix. For the raw count matrix, we used the Scanpy package in Python to preprocess the data. We used the command "scanpy.pp.normalize_total" to normalize counts to 10,000 per cell, and then logarithmized the data matrix using the command "sc.pp.log1p", followed by "scanpy.pp.scale" to scale data to unit variance

and zero mean. During the preprocessing process, we also provided options for users to decide if they need to filter highly variable genes or regress out mitochondrial genes. After preprocessing, we used the principal component analysis (PCA) (sc.tl.pca) for dimensionality reduction and logistic regression (sc.tl.rank_genes_groups(method= 'logreg')) to select cell type marker genes for each cell type. The final gene feature was the aggregation of the top n cell type marker genes. The preprocessing and marker gene selection step has been integrated into a single command called "find_marker_genes".

Pseudo-spots generation. The pseudo-spot was generated by randomly selecting a mixture of cells from scRNA-seq data. Based on our experience, the quality of the pseudo-spots is one of the key factors for the accuracy of the deconvolution results. A heterogeneous pseudo-spot pool is particularly important for the construction of a GCN link graph. Since the proportions of cell types are usually unbalanced in reference single cell data, regular pseudo-spot generation methods may generate an unbalanced pseudo-spot pool and less heterogeneity. To address this issue, we designed a sampling-by-cell-type method for obtaining a more balanced pseudo-spot generation. There are three hyper-parameters that need to be specified by the user for simulating the pseudo-spots: N_{up}, the upper limit of cell numbers in a pseudo-spot; N_{low}, the lower limit of cell numbers in a pseudo-spot; and K_{up}, the upper limit number of cell types in a pseudo-spot. The definition of K_{up} is that the number of cell types within a spot is usually limited in real ST datasets, which is different from bulk sequencing data. For each pseudo-spot simulation, we first randomly generated a number between 1 and K_{up}, representing the number of cell types (k) in a pseudo-spot, and then randomly selected k cell-types from the whole cell types (K_w) as the cell type set in the corresponding pseudo-spot. We next randomly selected a number (N_s) between N_{low} and N_{up} to determine the cell number in this pseudo-spot. We then randomly picked up N_s cells from the selected cell-type set to make up a pseudo-spot. The final expression level of the pseudo-spot was the summation of the counts of the randomly selected cells. The entire process of the pseudo-spot generation can be implemented using a single command "pseudo_spot_generation" in STdGCN.

Real-spots and pseudo-spots preprocessing and integration. The preprocessing of real-spots and pseudo-spots was similar to that applied to the reference single-cell data. After preprocessing, we only kept the previously selected marker genes for downstream analyses. For data integration, we employed PCA for dimensionality reduction to map real- and pseudo- spots into a common latent space. The integrated dimensionality reduction matrix was used as the input for the next link graph construction.

Link graph construction. The STdGCN system includes two link graphs, an expression graph and a spatial graph. The expression graph depicts the levels of expression similarity between spots. It is a hybrid graph consisting of three sub-graphs, a pseudo-spots internal graph, a real-spots internal graph, and a real-to-pseudo-spot graph. We employed the "cosine" distance to measure the distances between spots and the mutual nearest neighbors (MNN) [48] algorithm to search for links between spots. The MNN links between two spots (a and b) if and only if spot a belongs to one of the nearest neighbors of spot b, and spot b is also one of the nearest neighbors of spot a. For the real-to-pseudo-spots link graph, we

applied the integrated dimensionality reduction matrix of the two datasets as the input feature with the parameter "nearest neighbors = 20" per cell. For the two internal link graphs, data integration is not necessary, we thus directly used the normalized real-spot expression matrix and pseudo-spot expression matrix as the input features, respectively. The "nearest neighbors" were set as 10 for the real-spots internal graph and 20 for pseudo-spots internal graph. The adjacency matrix of the hybrid expression graph (A^{exp}) is a binary matrix. $A_{ab}^{exp} = 1$ represents a link between two spots and vice versa.

The spatial graph (A^{sp}) depicts the spatial distances between real-spots. We used the relative spatial coordinates to determine the Euclidean distance between one spot and all other spots (D_{ab}). To compute the relative spatial coordinates, we defined that the distance between two neighboring spots is a unit of distance (=1). For spatial graph construction, we only link spots with distances smaller than a pre-set (D_{upper}) radius. The adjacency matrix of the spatial graph is defined as follows:

$$A_{ab} = \begin{cases} 0 & when\ D_{ab} \geq D_{upper} \\ 1/D_{ab} & when\ D_{ab} < D_{upper} \end{cases}$$

Unlike the expression link graph being a binary matrix, the spatial graph is a weighted link graph based on the spatial distance between spots.

Model description. The STdGCN is built on the GCN [49], a semi-supervised learning algorithm that relies on graph-structured data. GCN is based on an efficient variant of convolutional neural networks which operate directly on graphs. Denoting V and E are the sets of nodes and edges of the graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, the objective of GCN is to learn the graph representations by combining embeddings of a graph node with its neighbors. The input of the model contains two parts, a feature matrix (X) representing the embeddings of all graph nodes and an adjacency matrix (A) describing the link graph structure. Each GCN layer can be written as a non-linear function

$$H^{(l+1)} = f\left(H^l, \tilde{A}\right) = \sigma(\tilde{A}H^l W^l)$$

where H^l is the output for the *l*-th layer with $H^l = X$, W^l is the weight matrix for the *l*-th layer that needs to be learned during the training, $\sigma(\cdot)$ is the activation function, and $\tilde{A} = D^{-1/2}(A/\lambda + I)D^{1/2}$ is the symmetric normalized adjacency matrix of A. The $\lambda = 20$ is a stable factor, as we found that adding this factor can increase the accuracy and robustness of STdGCN. I is the identity matrix and D is the diagonal degree matrix of $(A/\lambda + I)$.

The STdGCN builds a large and heterogeneous graph containing all pseudo-spots and real-spots as nodes so that the real-spots can be adapted from pseudo-spots. The objective of STdGCN is to predict the cell-type proportions of the real-spots. Therefore, the label for each spot is the cell-type proportions ($Y_{prop} \in \mathbb{R}^{(n_r+n_p)\times n_{type}}$), where n_r and n_p denote the real- and pseudo- spot numbers and n_{type} denote the cell-type number. We used the expression level for the real- and pseudo spots as the input matrix ($\mathrm{X} \in \mathbb{R}^{(n_r+n_p)\times g}$), where g represents the selected gene number. The input matrix was then forward propagated to two two-layer GCNs using the two previously constructed link graphs, an expression graph ($A^{exp} \in \mathbb{R}^{(n_r+n_p)\times(n_r+n_p)}$) and a spatial graph ($A^{sp} \in \mathbb{R}^{(n_r+n_p)\times(n_r+n_p)}$), with the function

$$\begin{cases} X_{exp} = ELU(\tilde{A}^{exp} ELU(\tilde{A}^{exp} X W_{exp}^{(0)}) W_{exp}^{(1)}) \\ X_{sp} = ELU(\tilde{A}^{sp} ELU(\tilde{A}^{sp} X W_{sp}^{(0)}) W_{sp}^{(1)}) \end{cases}$$

to obtain an expression embedding ($X_{exp} \in \mathbb{R}^{(n_r+n_p)\times m}$) and a spatial embedding ($X_{sp} \in \mathbb{R}^{(n_r+n_p)\times m}$). Where m is the dimension of the embedding space and $ELU(\cdot)$ is the activation function that can be expressed as

$$ELU(x) = \begin{cases} x, & if x > 0 \\ e^x - 1, & if x \leq 0 \end{cases}.$$

Note that to improve the robustness of the GCN model, we add a diagonal coefficient (λ) to the $\tilde{A}^{exp}$ and $\tilde{A}^{sp}$ to let $\tilde{A} = D^{-1/2}(A + \lambda I)D^{1/2}$, and found that $\lambda = 10$ exhibited a much better performance. We next concatenated the two GCN output embedding matrix ($X_{con} = [X_{exp}, X_{sp}] \in \mathbb{R}^{(n_r+n_p)\times 2m}$) as the input feature to forward propagate it to two another two-layer fully connected neural networks (FCNNs) with the function

$$\tilde{Y}_{prop} = softmax(b^{(1)} + a^1 ELU(b^{(0)} + a^{(0)} X_{con}))$$

where $a^{(l)}$ and $b^{(l)}$ indicate the weight matrix and additive bias for the *l*-th layer respectively, $\tilde{Y}_{prop} \in \mathbb{R}^{(n_r+n_p)\times n_{type}}$ is the prediction of cell-type proportions, and $softmax(\cdot)$ is the activation function that can be expressed as

$$softmax(x_i) = \frac{x_i}{\sum_j x_j}.$$

We used the KL divergence with L2 penalty as the loss function to train STdGCN.

Parameter settings. Herein, we used a two-layer GCN to process STdGCN, as we found that two layers performed better than one layer or more layers. We trained STdGCN for a maximum of 3,000 epochs using stochastic gradient descent (SGD) (Python command: torch.optim.SGD (model.parameters(), lr=0.2, momentum=0.9, weight_decay=0.0001, dampening=0, nesterov=True)) and stopped training if the validation loss did not decrease for 20 consecutive epochs. We used a self-adaption method orch.optim.lr_scheduler.ReduceLROnPlateau to adjust the decrease of the learning rate. The reduced learning rate factor was 0.1 and the number of epochs with no improvement after which learning rate would be reduced was set as 5.

Code Availability:

STdGCN and all script files used in the analysis in this manuscript can be downloaded from GitHub at https://github.com/luoyuanlab/stdgcn.

Data Availability:

All datasets used in this study are publicly available. The seqFISH dataset of the mouse embryos are available for download at https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/. The seqFISH+ dataset of the mouse mouse somatosensory region are available using the Giotto command getSpatialDataset(dataset = “seqfish_SS_cortex”, method = “wget”). The Slide-seq data of the mouse

testis are available for download at https://www.dropbox.com/s/ygzpj0d0oh67br0/Testis_Slideseq_Data.zip?dl=0. The MERFISH data of the mouse testis are available for download at https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248. The 10X Genomics Visium data of the human breast cancer are available for download at https://doi.org/10.5281/zenodo.4739739 and https://doi.org/10.5281/zenodo.3957257. The spatial transcriptomic data of the developing human heart are available for download at https://www.spatialresearch.org/resources-published-datasets/doi-10-1016-j-cell-2019-11-025/.

Table 1. The average ranks of the 14 benchmarked models across the four datasets.

	Model	seqFISH	seqFISH+	Slide-seq	MERFISH
JSD	STdGCN	1	1	1.33	1.82
	cell2location	7.17	3.5	8.83	13.07
	CellDART	8.17	7.5	8.33	4.8
	DestVI	3.5	14	14	10.87
	DSTG	14	12	13	13.8
	RCTD	2.17	4	3.67	1.82
	spatialDWLS	4.33	6	6.17	3.92
	SPOTlight	12.33	11.5	10.5	10.82
	STdeconvolve	11.17	12	11.17	6.81
	stereoscope	4.5	7	7	9.96
	STRIDE	12.5	10	8.83	8.57
	Tangram	9	9.5	6.83	6.55
	Adroit	5.5	3	3.5	2.96
	BayesPrism	9.67	4	1.83	9.24
RMSE	STdGCN	1	1	1.33	1.4
	cell2location	5.83	2.5	7.5	11.89
	CellDART	8.33	7.5	10.5	3.94
	DestVI	3	14	13.17	11.08
	DSTG	14	11.5	13.33	13.46
	RCTD	2.67	4	4.17	2.94
	spatialDWLS	6.5	6.5	6.17	4.96
	SPOTlight	12.83	11.5	9.5	9.94
	STdeconvolve	11.5	12.5	11.5	7.33
	stereoscope	4.33	7	7	12.45
	STRIDE	11.67	9	8.33	6.98
	Tangram	9.33	10.5	7.67	7.5
	Adroit	5	3	2.83	2.27
	BayesPrism	9	4.5	2	8.85

Note: JSD, Jensen–Shannon divergence; RMSE, root mean square error

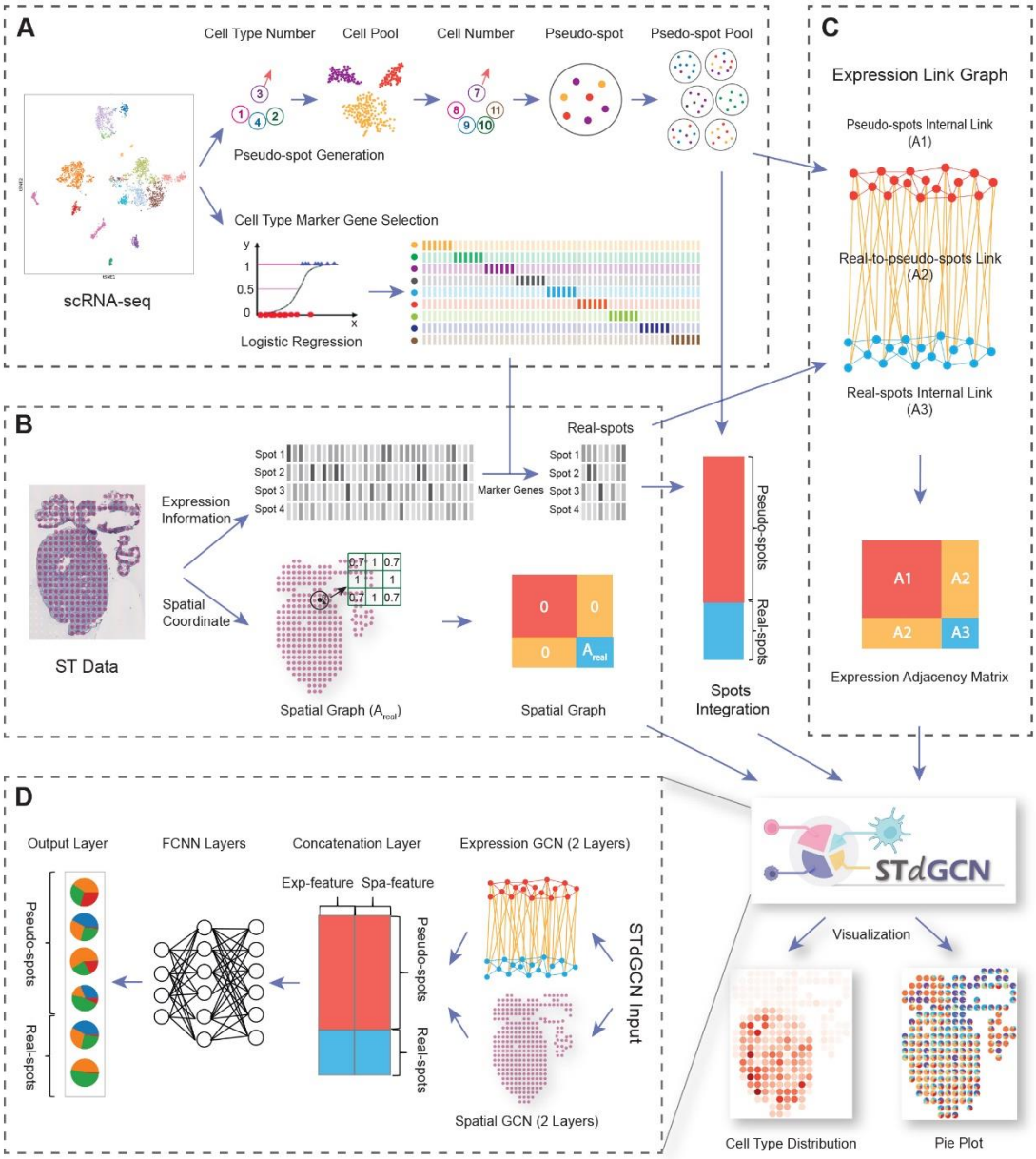

Figure 1 Schematic view of the STdGCN framework. STdGCN is designed for ST data cell type deconvolution by using scRNA-seq data as the reference. STdGCN first employs the scRNA-seq reference data to identify cell-type marker genes and simulate a pseudo-spot pool **(A)**. It then builds two link graphs, a spatial graph **(B)** and an expression graph **(C)**. The expression graph is a hybrid graph consisting of three sub-graphs, a pseudo-spot internal graph, a real-spot internal graph, and a real-to-pseudo-spot graph. Each subgraph is constructed by mutual nearest neighbors (MNN) based on their expression similarity. Based on the two link graphs, a GCN-based model is used to propagate both real- and pseudo- spots. The output of STdGCN is the cell-type proportions for each spot **(D)**. STdGCN can provide two types of figures to visualize the predicted results of the real-spot dataset.

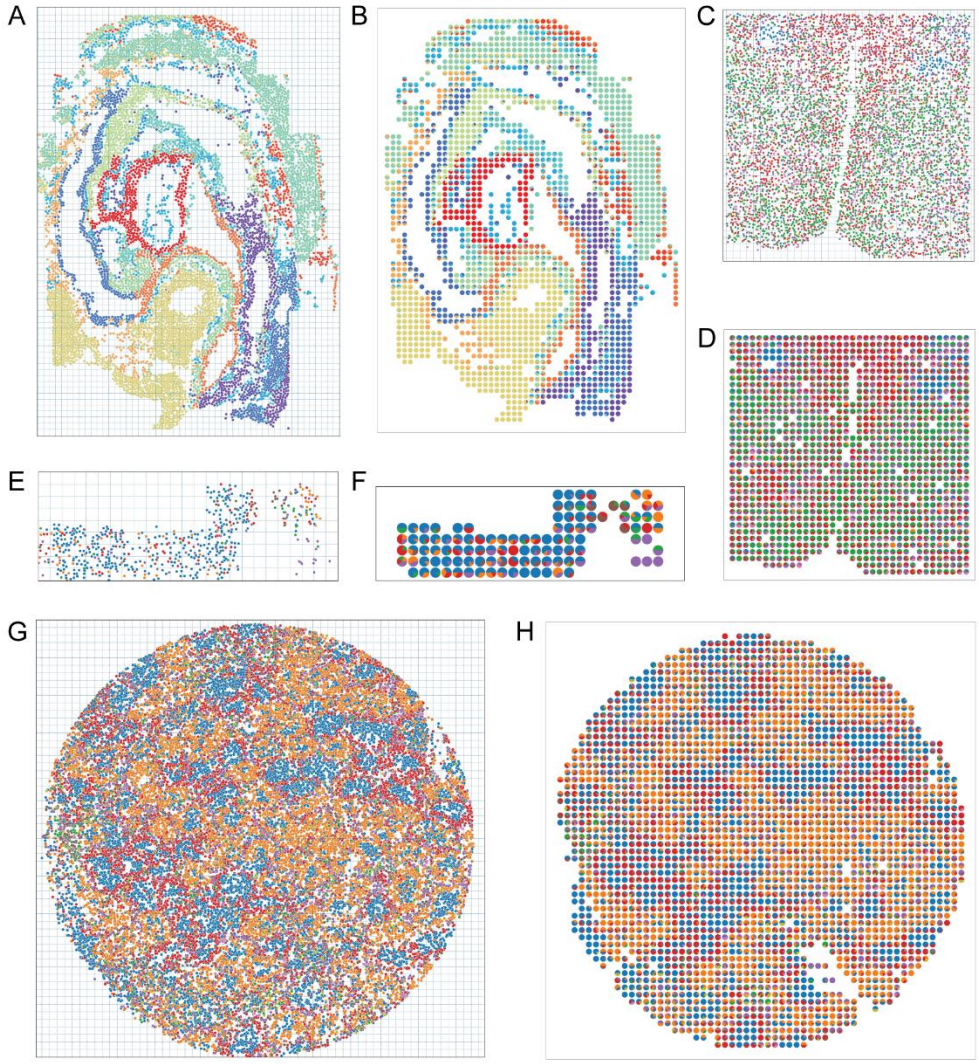

Figure 2 Four representative multi-cellular synthetic benchmark ST slices corresponding to the four ST platforms. The scatter plots display the cell atlas of the mouse embryos sequenced with seqFISH (**A**), mouse preoptic region sequenced with MERFISH (**C)**, mouse somatosensory (SS) region sequenced with seqFISH+ (**E**), and mouse testis region sequenced with Slide-seq (**G**). We resampled these datasets by splitting each slice into multiple square pixel areas (dash line). Cells within each square pixel area were merged into a synthetic spot. Pie plots display the cell-type proportions for each synthetic spot after resampling (**B, D, F, and H**).

Figure 3 Assessment of STdGCN and benchmarked methods for estimation of cell type proportions. (**A, B, C, and D**) To compare the performance of models with different numbers of cells within a spot, we split the spots into two groups: a smaller group (cell number in spot ≤5) and a larger group (cell number in spot >5). Bar plots display the Jensen-Shannon divergence (JSD) of the benchmarked methods for the two groups and the total spots across the four datasets. (**E, F, G, and H**) Comparison of the models in spots with different number of cell type mixtures. We split the spots into four groups based on the number of cell types within a spot: singleton (one cell type), doubleton (two cell types), tripleton (three cell types), and multiton (equal or greater than four cell types). Bar plots display the JSD of the benchmarked methods for the four groups and the total spots across the four datasets. (**I, J, K, and L**)

Bar plots display the JSD of STdGCN and benchmarked methods for distinguishing diverse cell types across the four datasets.

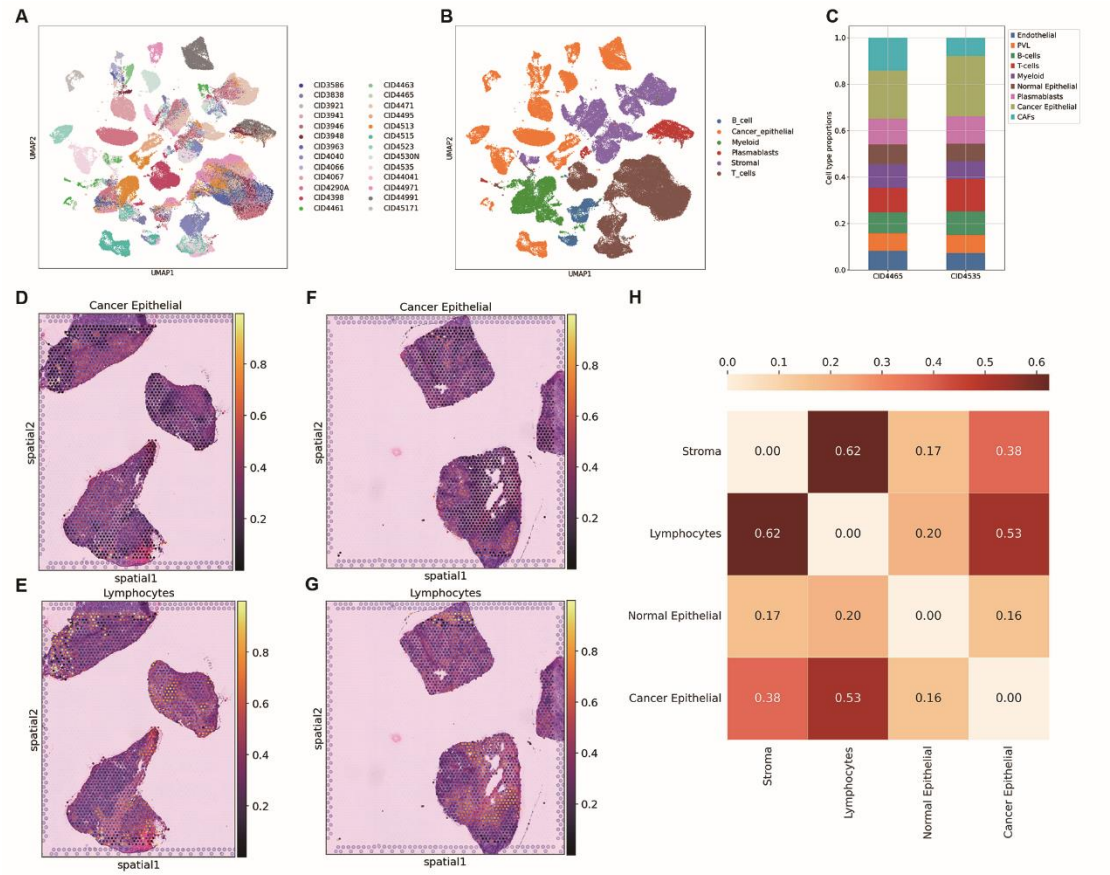

Figure 4 Performance of STdGCN on the human breast cancer 10X Genomics Visium dataset. (**A and B**) UMAP plots of the cells from the scRNA-seq dataset split by patient and cell type. (**C**) Mean predicted cell type proportions for the patient CID4465 and CID4535. (**D and E**) The predicted proportions of cancer cells and lymphocytes (integration of B-cells, T-cells, myeloid and plasmablasts) for the patient CID4535. (**F and G**) The predicted proportions of cancer cells and lymphocytes for the patient CID4465. (**H**) The heatmap displays the co-localization score (**Methods**) for normal cells, cancer cells, lymphocytes and stroma (integration of cancer-associated fibroblasts, perivascular-like and endothelial cells)

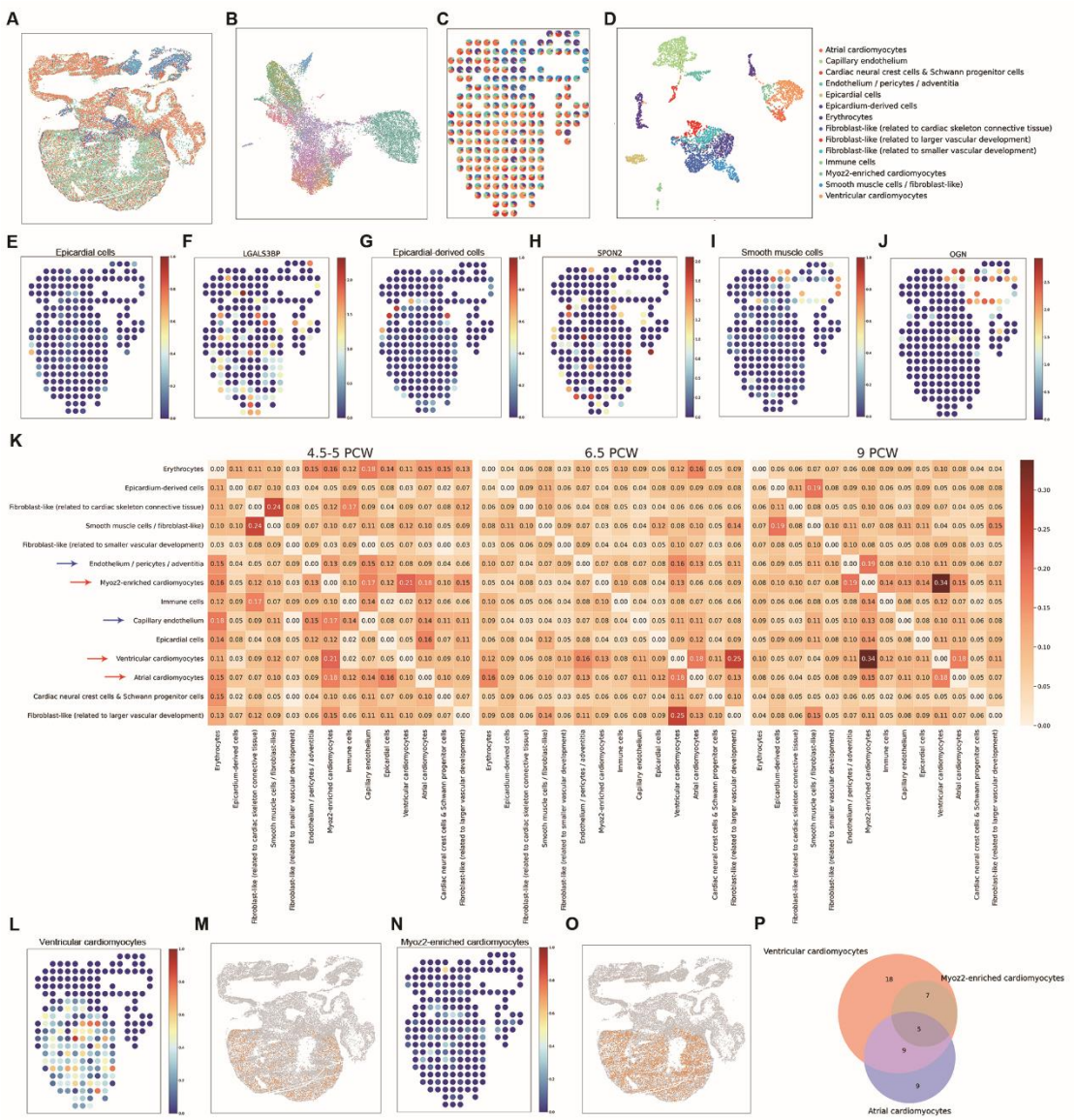

Figure 5 Performance of STdGCN on the developing human heart spatial transcriptomic dataset. (**A**) The spatial structure and cell type distribution of cells identified by in situ sequencing (ISS) in a PCW6.5 heart tissue section. (**B**) UMAP plot of the cells as in (**A**). (**C**) The pie plot displays the predicted cell-type proportions in each spot of the PCW6.5-6 heart tissue section. (**D**) UMAP plot of the cells from the scRNA-seq dataset. (**E, G, I, L, and N**) The scatter plot displays the predicted proportions of epicardium cells (**E**), epicardium-derived cells (**G**), smooth muscle cells / fibroblast-like cells (**I**), ventricular cardiomyocytes (**N**), and Myoz2-enriched cardiomyocytes (**P**) from the PCW6.5-6 heart tissue section. (**F, H, and J**) The expression levels of the corresponding cell-type marker genes. (**K**) The heatmap displays the co-localization score for all deconvolved cell types in the three stages (PCW4.5-5, PCW6.5, PCW9) of heart tissue sections. Red arrows are cardiomyocytes and blue arrows are endothelial cells. (**M**) The distributions of ventricular cardiomyocytes in the ISS dataset. (**O**) The distributions of Myoz2-enriched cardiomyocytes in the ISS dataset. (**P**) The Venn diagram displays the overlapping of cell-type marker genes between Ventricular cardiomyocytes, Myoz2-enriched cardiomyocytes, and Atrial Cardiomyocytes after filtering.