MERGE: Multi-faceted Hierarchical Graph-based GNN for Gene Expression Prediction from Whole Slide Histopathology Images

Supplementary Material

In the supplementary material, we begin with implementation details in Sec. 7. Sec. 7.1 discusses the experimental setup, including the hardware information and environment specifications. In Sec. 7.2 we discuss the nuances of the SPCS smoothing method and the details pertaining to its implementation for our purposes. We analyze the influence of cluster size on the feature space clustering and how well it aligns with the gene expressions in Sec. 7.3. Sec. 7.4 discusses the implementation details of our six baselines and also analyzes the performance gap in case of one of them. In the second section (Sec. 8), we further analyze the results of our experiments and present visualizations to that effect. Sec. 8.1 presents a comparison of the PCC scores obtained by MERGE and TRIPLEX [1] across the ST-Net dataset, while Sec. 8.2 presents visualizations for ablation of the graph construction strategy and its various modules.

7. Implementation Details

7.1. Experimental Setup

The ResNet18-based patch encoder is implemented in PyTorch (version 2.2.2). The graph neural network is implemented using PyTorch Geometric (version 2.5.2). Both models are trained on a NVIDIA RTX A6000 GPU. To ensure reproducibility a constant seed (3927) is set across all implementations and reruns of the model. The training of the ResNet18 is performed over 15 epochs, while the GNN is trained over 400 epochs. Each training process is replicated for five times and the best model is picked for each experiment.

7.2. Smoothing

We already established the effectiveness of SPCS [5] smoothing over the spatial smoothing employed in prior studies [1, 2]. This section presents further details on the implementation of the SPCS smoothing employed on the ST-Net dataset. The source code for SPCS is obtained from the GitHub repository provided in the original publication. The R-package is compiled and run in a Python kernel. The raw UMI counts and tissue position coordinates for the spots must first be converted to the appropriate data format to be processed by the SPCS code. In the original SPCS implementation, a quality check is performed across the gene library to filter out low-quality genes using a threshold for the percentage of spots in which a gene is expressed as well as the variance of the gene expressions. Typically the zero cutoff parameter is set to 0.7, meaning that genes that are zero in over 70% of the spots are discarded, while the rest are retained. Running this gene filtering step on all samples of the ST-Net dataset reveals that the vast majority of slides retain most of the genes (see Fig. 8). The average number of genes dropped per sample is 5.72. We skip this quality check step for all three datasets, and perform SPCS 250-gene subsets used in the TRIPLEX paper. These genes are listed in Fig. 9, Fig. 10, Fig. 11. The values for the other parameters in the SPCS method are outlined in Tab. 5.

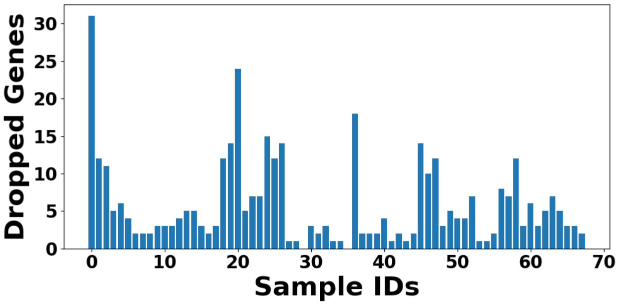


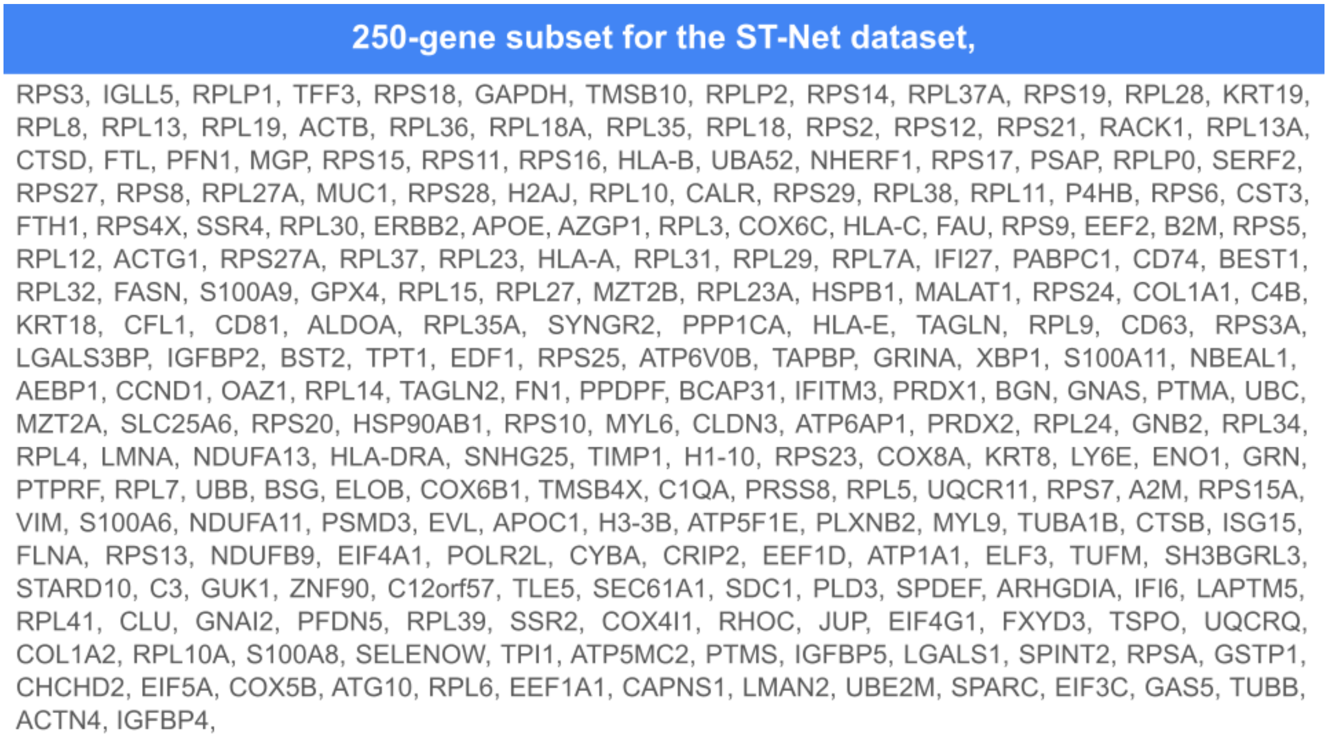
Figure 8. This presents a bar chart of the sample-by-sample dropping of genes based on amount of zero expression in the ST-Net dataset. The average number of genes dropped per sample is 5.72. We skip this step and use the 250-gene subset from the original study [2])

Parameter	Value
Gene Zero Cutoff	N/A
Gene Variance Cutoff	N/A
τ_s	2
τ_p	16
$alpha$	0.6
$beta$	0.4
Filling threshold	0.5
is.hexa	False
is.padding	False

Table 5. Parameters for running SPCS on the ST-Net dataset.

7.3. Cluster Size

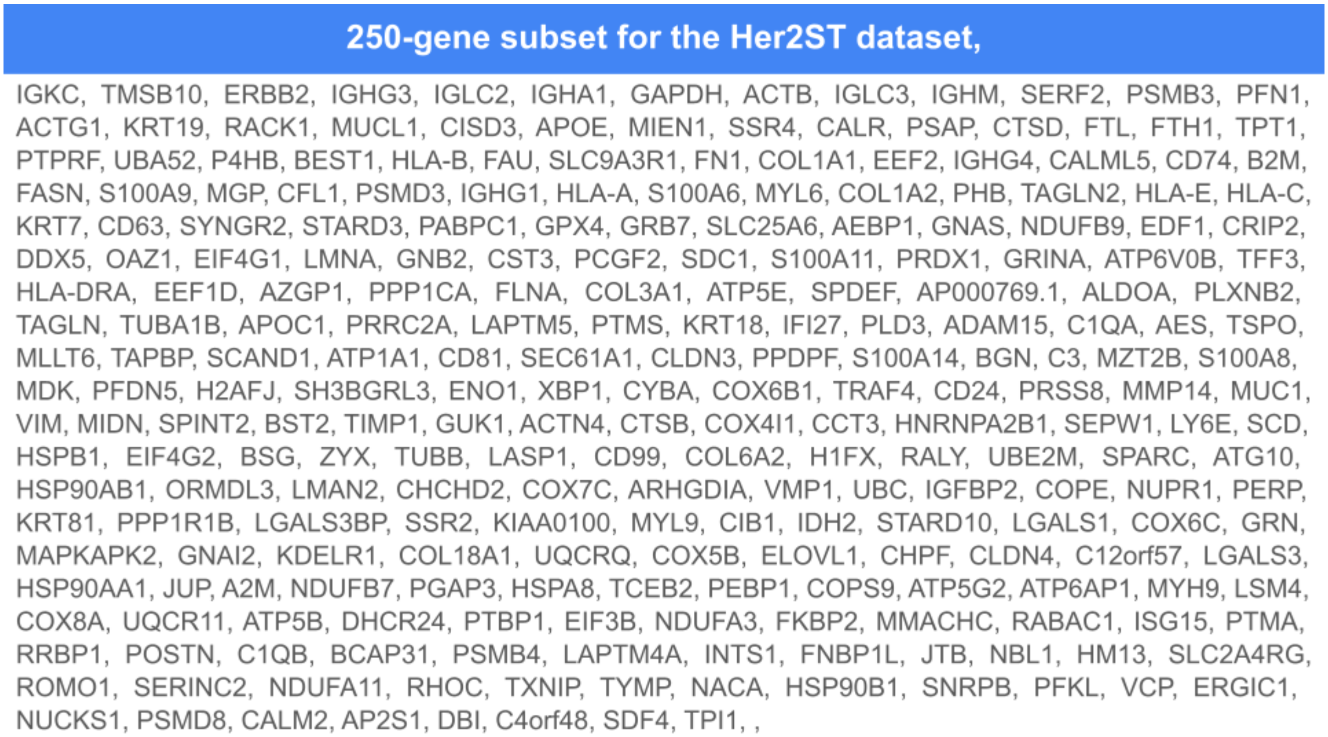
Spatial Clustering. The cluster size impacts the two clustering approaches in different ways. For the spatial clustering, using very small cluster sizes will mean that a smaller number of adjacent spots will be grouped together. This will result in the graph trying to model smaller tissue segments. This can have varying effects depending on the sample. If



250-gene subset for the ST-Net dataset,

RPS3, IGLL5, RPLP1, TFF3, RPS18, GAPDH, TMSB10, RPLP2, RPS14, RPL37A, RPS19, RPL28, KRT19, RPL8, RPL13, RPL19, ACTB, RPL36, RPL18A, RPL35, RPL18, RPS2, RPS12, RPS21, RACK1, RPL13A, CTSD, FTL, PFN1, MGP, RPS15, RPS11, RPS16, HLA-B, UBA52, NHERF1, RPS17, PSAP, RPLP0, SERF2, RPS27, RPS8, RPL27A, MUC1, RPS28, H2AJ, RPL10, CALR, RPS29, RPL38, RPL11, P4HB, RPS6, CST3, FTH1, RPS4X, SSR4, RPL30, ERBB2, APOE, AZGP1, RPL3, COX6C, HLA-C, FAU, RPS9, EEF2, B2M, RPS5, RPL12, ACTG1, RPS27A, RPL37, RPL23, HLA-A, RPL31, RPL29, RPL7A, IFI27, PABPC1, CD74, BEST1, RPL32, FASN, S100A9, GPX4, RPL15, RPL27, MZT2B, RPL23A, HSPB1, MALAT1, RPS24, COL1A1, C4B, KRT18, CFL1, CD81, ALDOA, RPL35A, SYNGR2, PPP1CA, HLA-E, TAGLN, RPL9, CD63, RPS3A, LGALS3BP, IGFBP2, BST2, TPT1, EDF1, RPS25, ATP6V0B, TAPBP, GRINA, XBP1, S100A11, NBEAL1, AEBP1, CCND1, OAZ1, RPL14, TAGLN2, FN1, PPDPF, BCAP31, IFITM3, PRDX1, BGN, GNAS, PTMA, UBC, MZT2A, SLC25A6, RPS20, HSP90AB1, RPS10, MYL6, CLDN3, ATP6AP1, PRDX2, RPL24, GNB2, RPL34, RPL4, LMNA, NDUFA13, HLA-DRA, SNHG25, TIMP1, H1-10, RPS23, COX8A, KRT8, LY6E, ENO1, GRN, PTPRF, RPL7, UBB, BSG, ELOB, COX6B1, TMSB4X, C1QA, PRSS8, RPL5, UQCR11, RPS7, A2M, RPS15A, VIM, S100A6, NDUFA11, PSMD3, EVL, APOC1, H3-3B, ATP5F1E, PLXNB2, MYL9, TUBA1B, CTSB, ISG15, FLNA, RPS13, NDUFB9, EIF4A1, POLR2L, CYBA, CRIP2, EEF1D, ATP1A1, ELF3, TUFM, SH3BGRL3, STARD10, C3, GUK1, ZNF90, C12orf57, TLE5, SEC61A1, SDC1, PLD3, SPDEF, ARHGDIA, IFI6, LAPTM5, RPL41, CLU, GNAI2, PFDN5, RPL39, SSR2, COX4I1, RHOC, JUP, EIF4G1, FXYD3, TSPO, UQCRQ, COL1A2, RPL10A, S100A8, SELENOW, TPI1, ATP5MC2, PTMS, IGFBP5, LGALS1, SPINT2, RPSA, GSTP1, CHCHD2, EIF5A, COX5B, ATG10, RPL6, EEF1A1, CAPNS1, LMAN2, UBE2M, SPARC, EIF3C, GAS5, TUBB, ACTN4, IGFBP4,

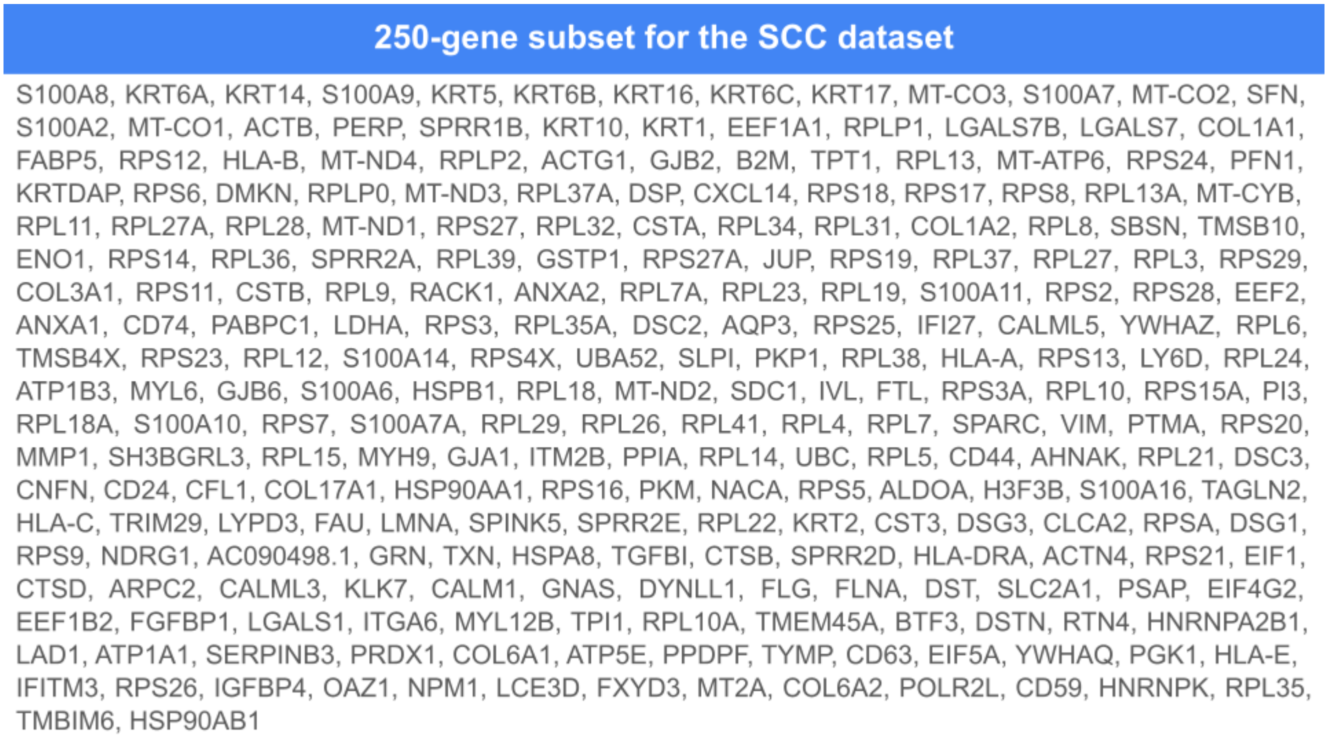
Figure 9. List of the 250 genes used in all experiments on the ST-Net dataset.



250-gene subset for the Her2ST dataset,

IGKC, TMSB10, ERBB2, IGHG3, IGLC2, IGHA1, GAPDH, ACTB, IGLC3, IGHM, SERF2, PSMB3, PFN1, ACTG1, KRT19, RACK1, MUCL1, CISD3, APOE, MIEN1, SSR4, CALR, PSAP, CTSD, FTL, FTH1, TPT1, PTPRF, UBA52, P4HB, BEST1, HLA-B, FAU, SLC9A3R1, FN1, COL1A1, EEF2, IGHG4, CALML5, CD74, B2M, FASN, S100A9, MGP, CFL1, PSMD3, IGHG1, HLA-A, S100A6, MYL6, COL1A2, PHB, TAGLN2, HLA-E, HLA-C, KRT7, CD63, SYNGR2, STARD3, PABPC1, GPX4, GRB7, SLC25A6, AEBP1, GNAS, NDUFB9, EDF1, CRIP2, DDX5, OAZ1, EIF4G1, LMNA, GNB2, CST3, PCGF2, SDC1, S100A11, PRDX1, GRINA, ATP6V0B, TFF3, HLA-DRA, EEF1D, AZGP1, PPP1CA, FLNA, COL3A1, ATP5E, SPDEF, AP000769.1, ALDOA, PLXNB2, TAGLN, TUBA1B, APOC1, PRRC2A, LAPTM5, PTMS, KRT18, IFI27, PLD3, ADAM15, C1QA, AES, TSPO, MLLT6, TAPBP, SCAND1, ATP1A1, CD81, SEC61A1, CLDN3, PPDPF, S100A14, BGN, C3, MZT2B, S100A8, MDK, PFDN5, H2AFJ, SH3BGRL3, ENO1, XBP1, CYBA, COX6B1, TRAF4, CD24, PRSS8, MMP14, MUC1, VIM, MIDN, SPINT2, BST2, TIMP1, GUK1, ACTN4, CTSB, COX4I1, CCT3, HNRNPA2B1, SEPW1, LY6E, SCD, HSPB1, EIF4G2, BSG, ZYX, TUBB, LASP1, CD99, COL6A2, H1FX, RALY, UBE2M, SPARC, ATG10, HSP90AB1, ORMDL3, LMAN2, CHCHD2, COX7C, ARHGDIA, VMP1, UBC, IGFBP2, COPE, NUPR1, PERP, KRT81, PPP1R1B, LGALS3BP, SSR2, KIAA0100, MYL9, CIB1, IDH2, STARD10, LGALS1, COX6C, GRN, MAPKAPK2, GNAI2, KDELR1, COL18A1, UQCRQ, COX5B, ELOVL1, CHPF, CLDN4, C12orf57, LGALS3, HSP90AA1, JUP, A2M, NDUFB7, PGAP3, HSPA8, TCEB2, PEBP1, COPS9, ATP5G2, ATP6AP1, MYH9, LSM4, COX8A, UQCR11, ATP5B, DHCR24, PTBP1, EIF3B, NDUFA3, FKBP2, MMACHC, RABAC1, ISG15, PTMA, RRBP1, POSTN, C1QB, BCAP31, PSMB4, LAPTM4A, INTS1, FNBP1L, JTB, NBL1, HM13, SLC2A4RG, ROMO1, SERINC2, NDUFA11, RHOC, TXNIP, TYMP, NACA, HSP90B1, SNRPB, PFKL, VCP, ERGIC1, NUCKS1, PSMD8, CALM2, AP2S1, DBI, C4orf48, SDF4, TPI1, ,

Figure 10. List of the 250 genes used in all experiments on the Her2ST dataset.



250-gene subset for the SCC dataset

S100A8, KRT6A, KRT14, S100A9, KRT5, KRT6B, KRT16, KRT6C, KRT17, MT-CO3, S100A7, MT-CO2, SFN, S100A2, MT-CO1, ACTB, PERP, SPRR1B, KRT10, KRT1, EEF1A1, RPLP1, LGALS7B, LGALS7, COL1A1, FABP5, RPS12, HLA-B, MT-ND4, RPLP2, ACTG1, GJB2, B2M, TPT1, RPL13, MT-ATP6, RPS24, PFN1, KRTDAP, RPS6, DMKN, RPLP0, MT-ND3, RPL37A, DSP, CXCL14, RPS18, RPS17, RPS8, RPL13A, MT-CYB, RPL11, RPL27A, RPL28, MT-ND1, RPS27, RPL32, CSTA, RPL34, RPL31, COL1A2, RPL8, SBSN, TMSB10, ENO1, RPS14, RPL36, SPRR2A, RPL39, GSTP1, RPS27A, JUP, RPS19, RPL37, RPL27, RPL3, RPS29, COL3A1, RPS11, CSTB, RPL9, RACK1, ANXA2, RPL7A, RPL23, RPL19, S100A11, RPS2, RPS28, EEF2, ANXA1, CD74, PABPC1, LDHA, RPS3, RPL35A, DSC2, AQP3, RPS25, IFI27, CALML5, YWHAZ, RPL6, TMSB4X, RPS23, RPL12, S100A14, RPS4X, UBA52, SLPI, PKP1, RPL38, HLA-A, RPS13, LY6D, RPL24, ATP1B3, MYL6, GJB6, S100A6, HSPB1, RPL18, MT-ND2, SDC1, IVL, FTL, RPS3A, RPL10, RPS15A, PI3, RPL18A, S100A10, RPS7, S100A7A, RPL29, RPL26, RPL41, RPL4, RPL7, SPARC, VIM, PTMA, RPS20, MMP1, SH3BGRL3, RPL15, MYH9, GJA1, ITM2B, PPIA, RPL14, UBC, RPL5, CD44, AHNAK, RPL21, DSC3, CNFN, CD24, CFL1, COL17A1, HSP90AA1, RPS16, PKM, NACA, RPS5, ALDOA, H3F3B, S100A16, TAGLN2, HLA-C, TRIM29, LYPD3, FAU, LMNA, SPINK5, SPRR2E, RPL22, KRT2, CST3, DSG3, CLCA2, RPSA, DSG1, RPS9, NDRG1, AC090498.1, GRN, TXN, HSPA8, TGFBI, CTSB, SPRR2D, HLA-DRA, ACTN4, RPS21, EIF1, CTSD, ARPC2, CALML3, KLK7, CALM1, GNAS, DYNLL1, FLG, FLNA, DST, SLC2A1, PSAP, EIF4G2, EEF1B2, FGFBP1, LGALS1, ITGA6, MYL12B, TPI1, RPL10A, TMEM45A, BTF3, DSTN, RTN4, HNRNPA2B1, LAD1, ATP1A1, SERPINB3, PRDX1, COL6A1, ATP5E, PPDPF, TYMP, CD63, EIF5A, YWHAQ, PGK1, HLA-E, IFITM3, RPS26, IGFBP4, OAZ1, NPM1, LCE3D, FXYD3, MT2A, COL6A2, POLR2L, CD59, HNRNPK, RPL35, TMBIM6, HSP90AB1

Figure 11. List of the 250 genes used in all experiments on the SCC dataset.

a sample consists of large homogeneous tissue segments, smaller spatial clusters will not be too helpful in capturing the interactions of spots within them. But if there are small homogeneous tissue segments, using smaller clusters will help the GNN learn from these smaller groups of spots. This will result in a more accurate modeling of the morphology driven interactions among spots, and thereby enhance gene expression prediction.

Feature Space Clustering. In the feature space, the goal is to group spots based on imaging features. The expectation is that morphologically similar spots will have similar imaging features, and therefore, be grouped together. This extends to the idea that morphologically similar tissue segments are likely to exhibit similar gene expressions. Therefore, we should expect to see some alignment among tissue morphology, feature space clusters and the gene expressions. To visualize this we perform clustering in the gene space by using the gene expression vectors as the feature vectors in a clustering scheme similar to feature space clustering. The genes used for this clustering are - FASN, GNAS, XBP1, AEBP1, SPARC, and BGN [2]. These are all known cancer biomarkers. We use the same cluster size for both. Fig. 12 shows the outputs of feature space clustering and gene-based clustering for various cluster sizes. We can see that smaller clusters fail to capture morphologically meaningful groups well. They also fail to align well with gene space clusters. But larger cluster sizes result in more gene-aligned feature space clusters. We can see that there is still a significant misalignment among the clusters in the image feature space and the clusters in the gene expression space. This is why feature space clustering is not sufficient on its own, and provides better outputs only when combined with spatial clustering.

7.4. Baselines

7.4.1 ResNet+FCN

The first baseline is composed of our ResNet18 patch encoder followed by a fully connected layer (FCN) to predict the 250 genes per patch. This is the simplest variant of architecture, which is directly inspired by the original ST-Net [2] architecture where a DenseNet followed by a fully connected layer was used for gene expression prediction.

7.4.2 BLEEP

The source code of BLEEP [7] is obtained from the original publication. The same ResNet18 architecture is used here as the patch encoders in both MERGE and TRIPLEX. BLEEP seems to perform rather poorly when trained and evaluated on the SPCS smoothed data. Our assumption is that this is caused by the use of Harmony [4] by BLEEP. Harmony is a batch correction algorithm designed for sin-

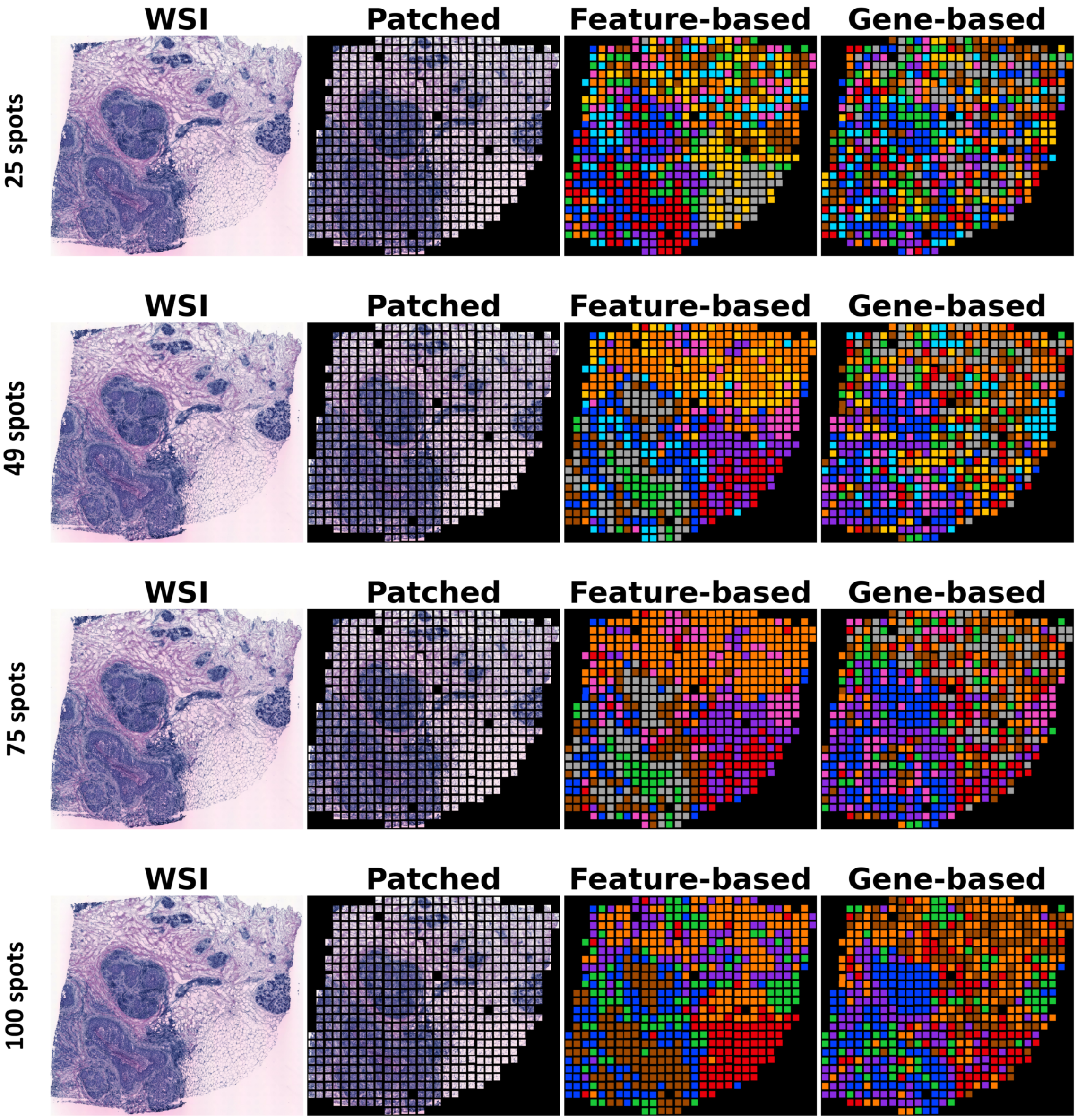


Figure 12. The figure shows the WSI, extracted patches, outputs of feature space clustering, and outputs of gene space clustering using five cancer biomarkers. Each row represents a different cluster size. It is evident that clusters that are too small fail to represent both the morphologically similar tissue segments and the gene space clusters. Larger cluster sizes are more effective at capturing both of these, but still can not accurately align with the gene space clusters. This underlines the necessity of feature space clustering but also depicts why it is not sufficient by itself.

gle cell RNA-seq data (scRNA-seq). Harmony models and removes the artifacts generated by known sources of variation in scRNA-seq data. In applying Harmony to ST data however, each spot must be treated as a cell and this does not translate well in all cases. In case of BLEEP, using Harmony on the raw gene expression data works well with their original four-sample dataset. However, using it on the SPCS smoothed ST-Net data interferes with the gene expression patterns captured by SPCS. We assume that this is what most likely results in a poor performance on the ST-Net dataset. The detrimental effect of applying Harmony on ST-Net data is visualized in Fig. 13 where we can see that the morphological patterns captured by SPCS for two cancer biomarker genes - FASN and GNAS - are lost in the outputs of Harmony.

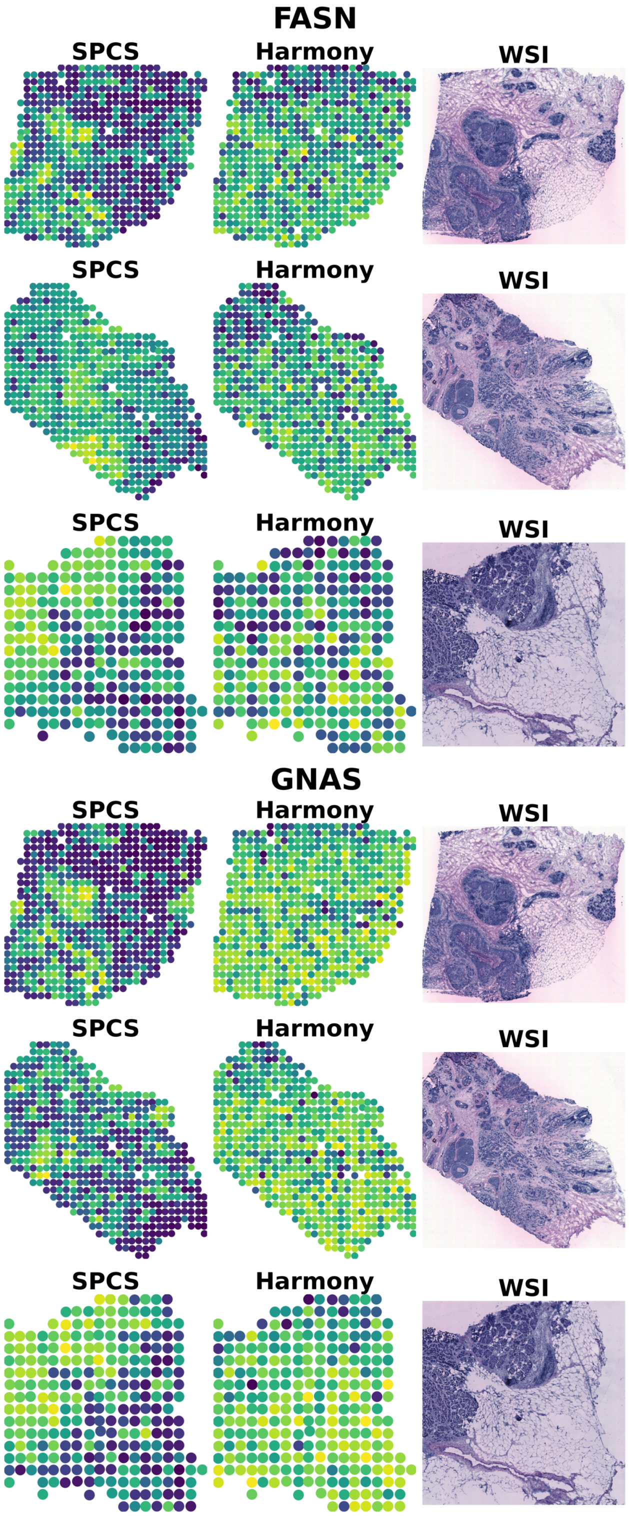


Figure 13. When we plot the expression values for two cancer biomarker genes - FASN and GNAS - on the tissue space, we see that the morphological patterns captured by SPCS are lost upon applying Harmony. This is most likely the cause of the poor performance of BLEEP on the SPCS smoothed ST-Net dataset.

7.4.3 HisToGene, Hist2ST and THItoGene

The source code of HisToGene [6], Hist2ST [8] and THItoGene [3] are adopted from the original publications. We train and test on SPCS smoothed data for all three baselines. All three models perform logCPM normalization on the gene expression matrix. We remove this portion of the code before training on SPCS data as we already perform logCPM normalization during SPCS smoothing. Hist2ST and THItoGene use 4-nearest-neighbors graphs for their purposes and we adapt that to the other datasets as well. We set the global seed for all modules and libraries the same value as MERGE (3927) for consistency.

7.4.4 TRIPLEX

The source code of TRIPLEX [1] is obtained from the original publication as well. This is used to train and test on both 8n and SPCS smoothed data. Since TRIPLEX performs the spatial smoothing (8n) within the model code, we remove that portion of code as well as the log normalization step while training on the SPCS smoothed data. The remaining parameters are kept unchanged in the original implementation and the parameter specifications are obtained from the supplementary materials provided by the authors. The seed set by TRIPLEX is also updated to match the seed in MERGE (3927) for consistency.

8. Results Analysis

8.1. Results Comparison - PCC

This section discusses the PCC scores attained by MERGE and TRIPLEX across the samples in ST-Net dataset for two cancer-relevant genes - FASN and GNAS. Fig. 14 shows two bar charts depicting the PCC attained by MERGE and TRIPLEX for each sample in ST-Net. The legend mentions the average PCC over the dataset for each method. It is evident that MERGE achieves a higher average PCC for both the FASN (0.39) gene and the GNAS (0.42) gene. Additionally, we can see that there are plenty of samples where MERGE achieves a higher PCC for both genes while TRIPLEX is unable to do so. The reverse however is rarely true. In case of FASN, MERGE performs better than TRIPLEX in *39 samples*, with an average PCC that is higher by 0.33. In case of GNAS, in the *41 samples* where MERGE performs better than TRIPLEX, the average PCC achieved by MERGE is higher that that of TRIPLEX by 0.34. Fig. 15 shows the sample-wise bar charts of the PCC achieved by both methods for the two genes. For visual convenience, the vertical red dashed line in each panel splits the chart into two zones. The zone on the left can be considered a *low PCC zone* where a method has achieved a low PCC, less than 0.25 for FASN and less than 0.265 for GNAS. We can see that the number of samples for this region of the

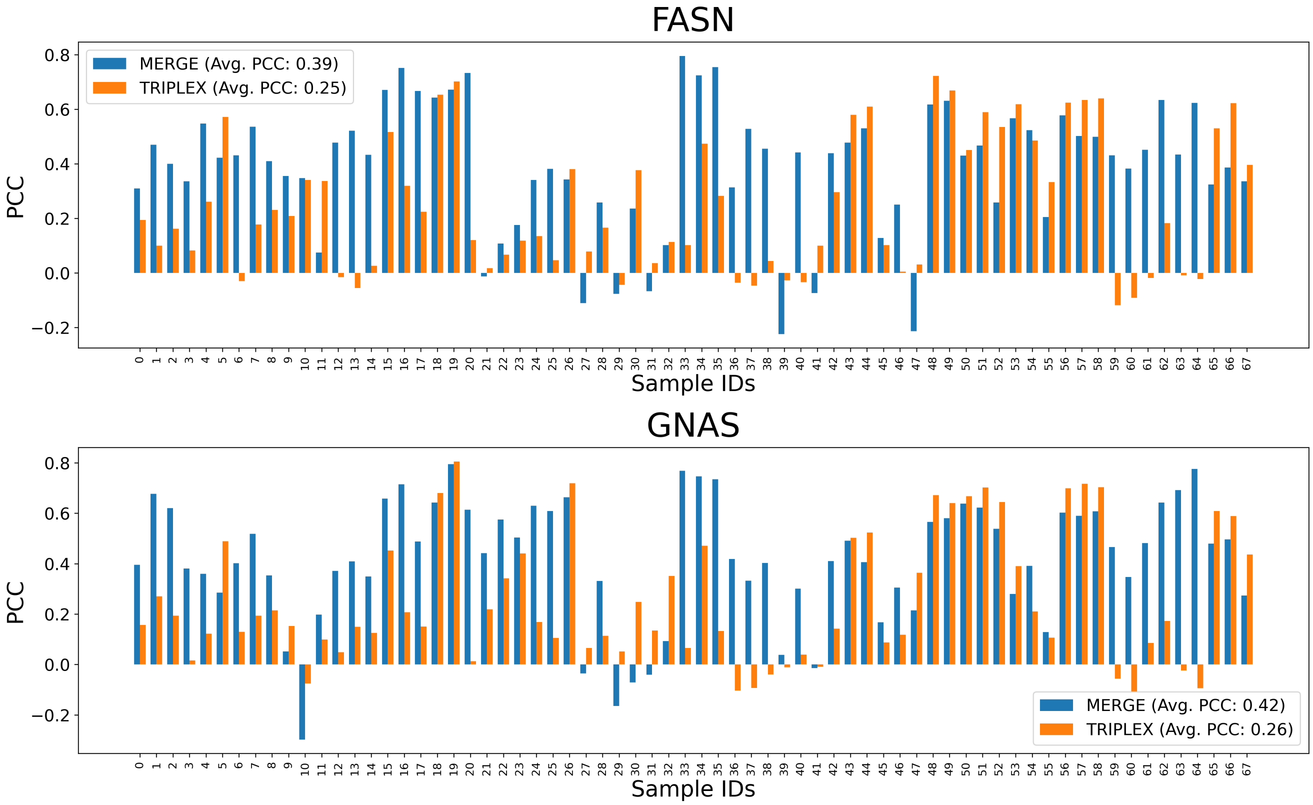


Figure 14. The PCC between the ground truth and predicted gene expressions for the tumor marker FNAS gene and breast cancer biomarker GNAS gene. The average PCC across the dataset is reported inside the legend within parentheses.

graph is significantly lower for MERGE (blue bars), which means that MERGE exhibits much fewer low PCC samples.

8.2. Ablation Visualization

Using the same ResNet18 based encoder and progressively adding the three modules of our graph construction strategy, we can see a comparative performance of the modules across multiple samples. When we plot the WSI and the ground truth gene expressions alongside the predicted expressions using one-hop edges, feature space clustering-based edges, spatial clustering-based edges, and the combined strategy (MERGE) - we can see that the Pearson Correlation Coefficient of predicted and ground truth gene expressions improves progressively. Fig. 16 and Fig. 17 visualize this comparison across multiple samples for the tumor marker FNAS gene. Similarly, Fig. 18 and Fig. 19 visualize this comparison across multiple samples for the breast cancer biomarker GNAS gene. For most samples, the PCC achieved using only feature space or spatial clustering is better than that achieved using only one-hop edges. The PCC is highest when using a combination of both clustering methods alongside the one-hop edges.

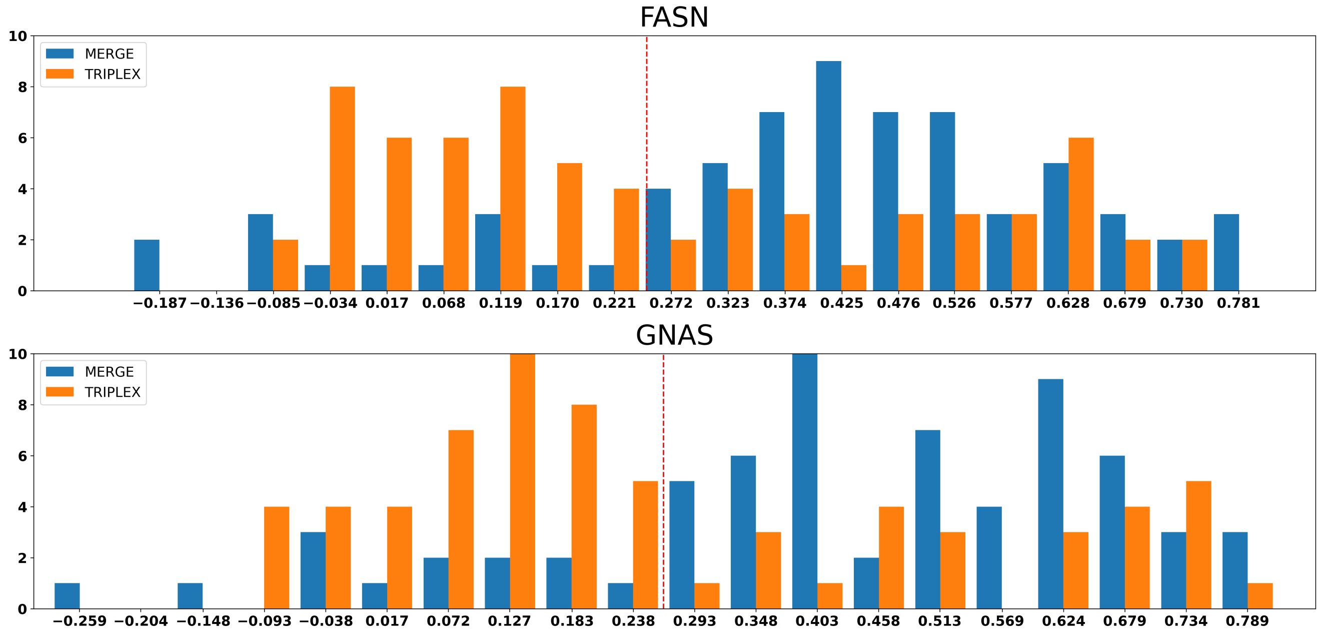


Figure 15. This is a histogram of PCC for the two methods for the ST-Net dataset. The upper panel shows the histogram for the FASN gene, while the lower panel shows the same for the GNAS gene. We can see that MERGE attains low PCC for significantly fewer samples. Additionally, there are very few samples where both methods perform poorly and attain negative PCC for either gene, although this number is slightly higher for MERGE.

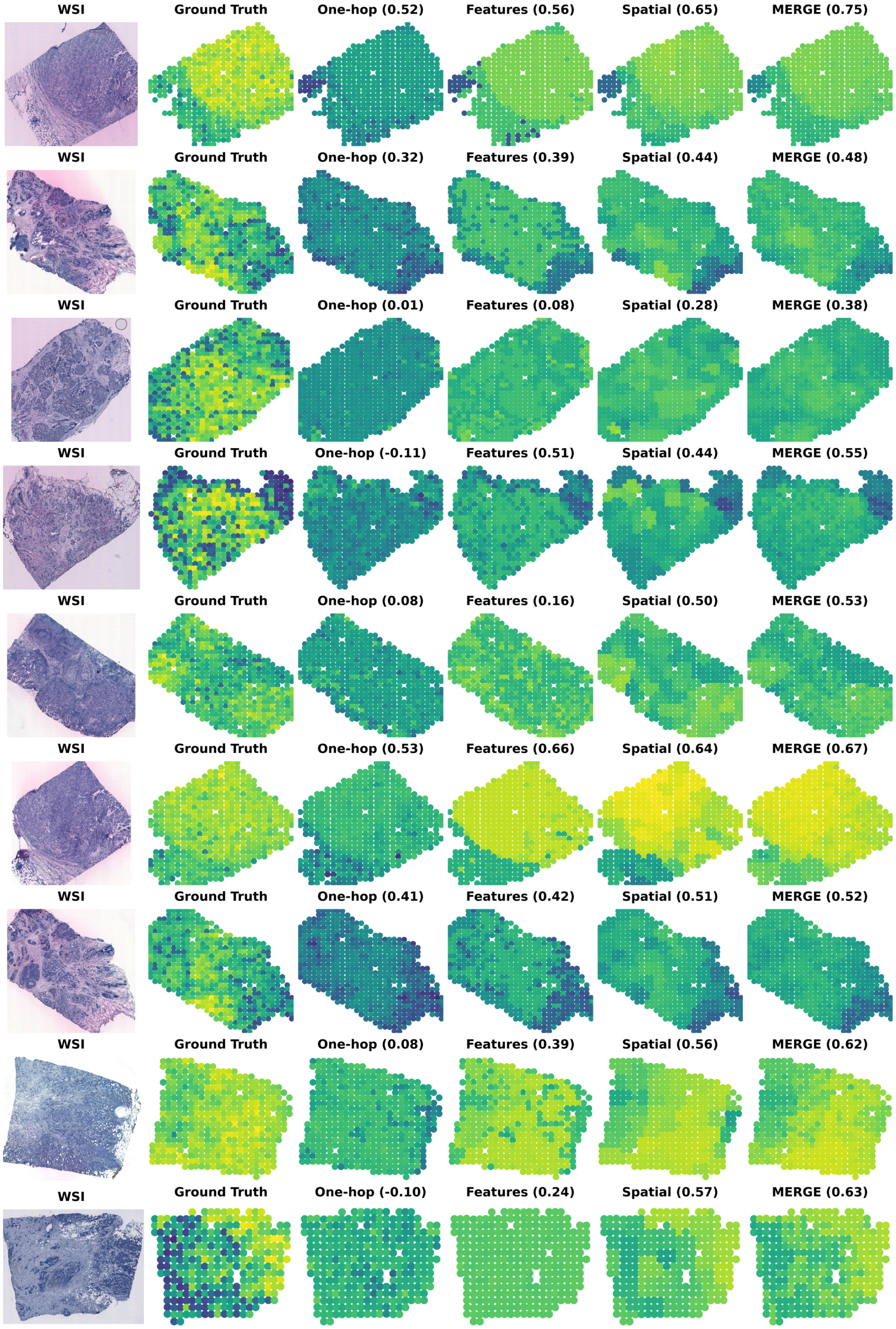


Figure 16. Figure shows PCC between ground truth expressions and predictions for the gene FASN in a few samples. Each row represents a sample, and from the left we have the WSI, the ground truth expressions, predicted expressions using one-hop edges, feature space clustering based edges, spatial clustering based edges and both clustering methods (MERGE).

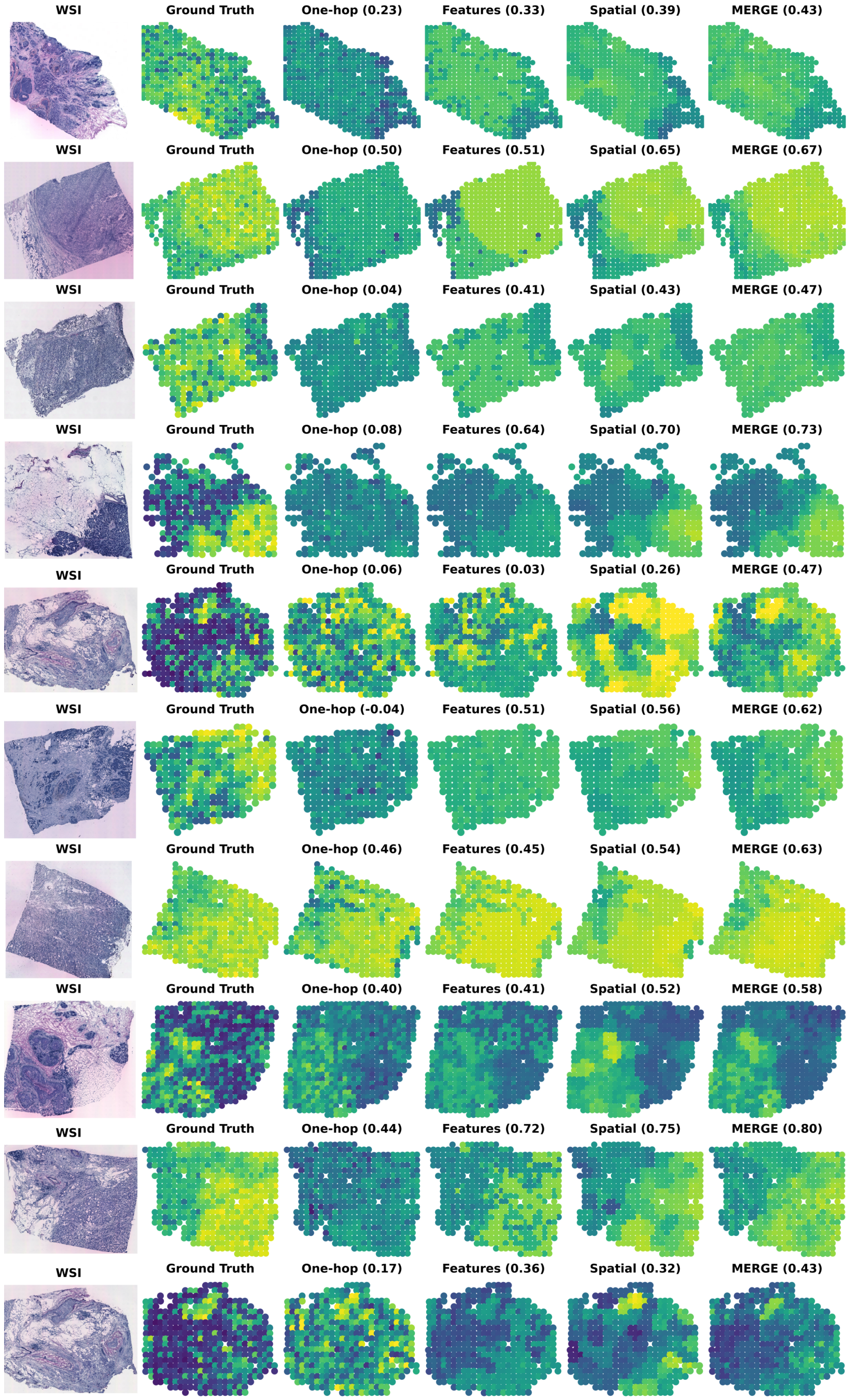


Figure 17. Extension of Fig. 16

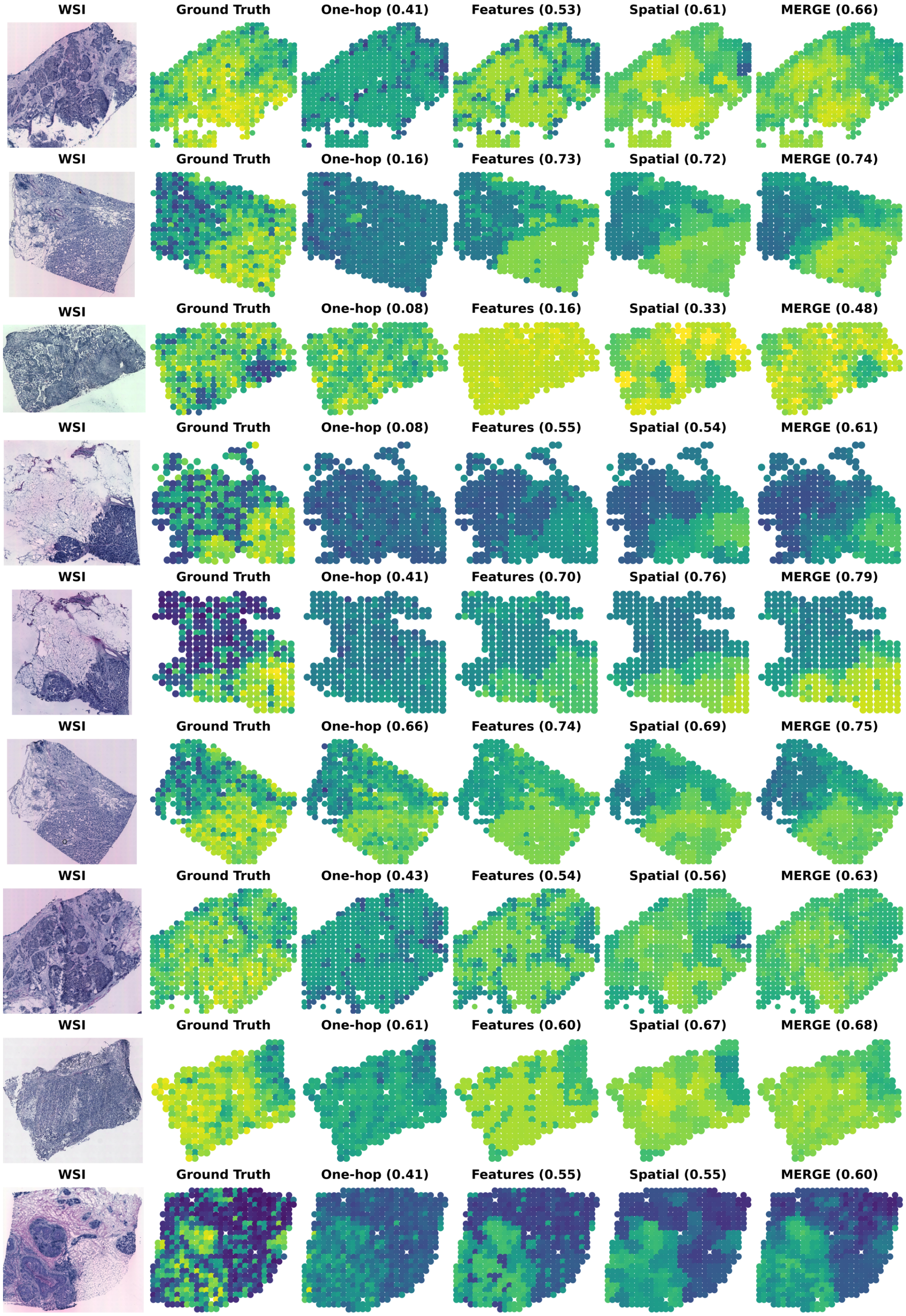


Figure 18. Figure shows PCC between ground truth expressions and predictions for the gene GNAS in a few samples. Each row represents a sample, and from the left we have the WSI, the ground truth expressions, predicted expressions using one-hop edges, feature space clustering based edges, spatial clustering based edges and both clustering methods (MERGE).

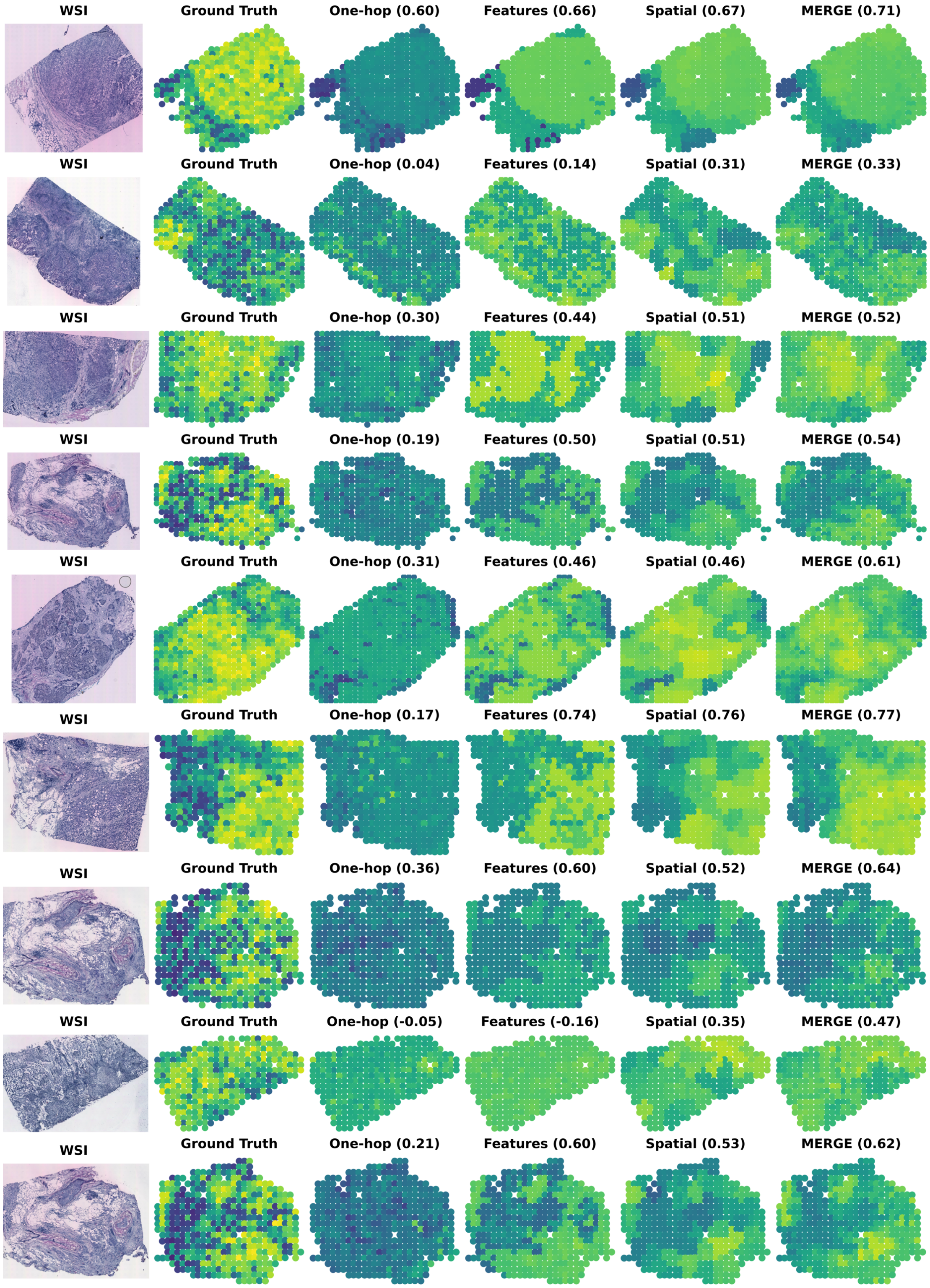


Figure 19. Extension of Fig. 18

References

[1] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024.

[2] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.

[3] Yuran Jia, Junliang Liu, Li Chen, et al. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1): bbad464, 2024.

[4] Ilya Korsunsky, Nghia Millard, Jean Fan, et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[5] Yusong Liu, Tongxin Wang, Ben Duggan, et al. SPCS: a spatial and pattern combined smoothing method for spatial transcriptomic expression. *Briefings in Bioinformatics*, 23(3): bbac116, 2022.

[6] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.

[7] Ronald Xie, Kuan Pang, Sai W Chung, et al. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, et al. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.