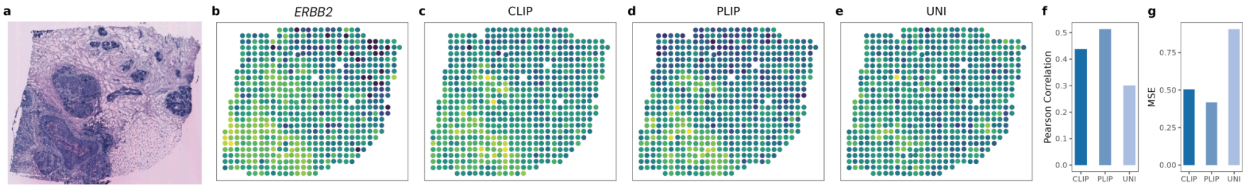

Figure R1 Gene expression prediction evaluation. (a) H2 histology images. (b) ERBB2 gene expression. (c) Predicted ERBB2 expression using zero-shot CLIP image embedding. (d) Predicted ERBB2 expression using zero-shot PLIP image embedding.(e) Predicted ERBB2 expression using zero-shot UNI image embedding. (f) Pearson Correlation of three models. (g) MSE of three models.

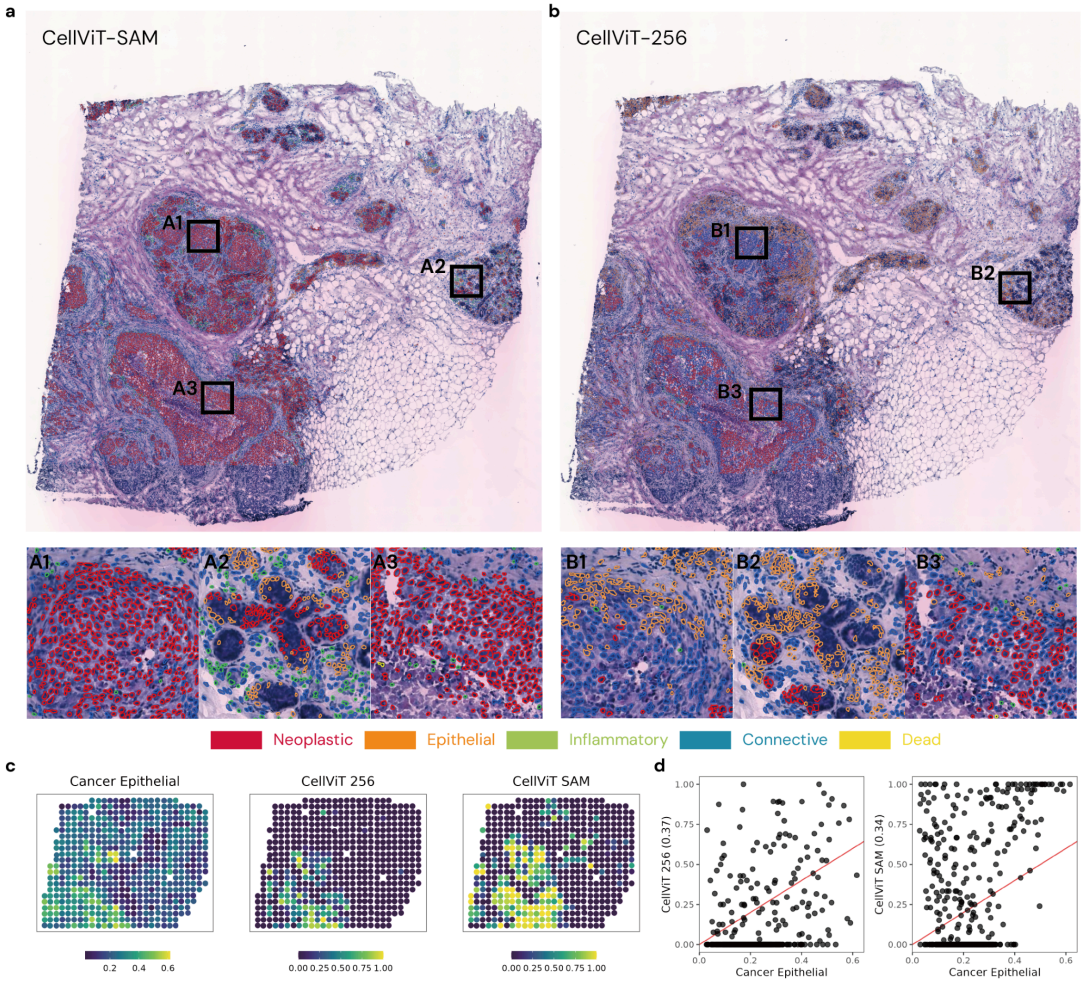

Figure R2 Cell segmentation evaluation. (a) CellViT-SAM model segmented cells. (b) CellViT-256 model segmented cells. (c) left: RCTD inferred cancer epithelial cell proportion in spot leve., middle: aggregated CellViT-256 segmented neoplastic cell proportion. Right: aggregated CellViT-SAM segmented neoplastic cell proportion. (d) scatter plot of RCTD inferred cell type proportion and aggregated segmented cell proportion. Ylab: model name (pearson correlation).

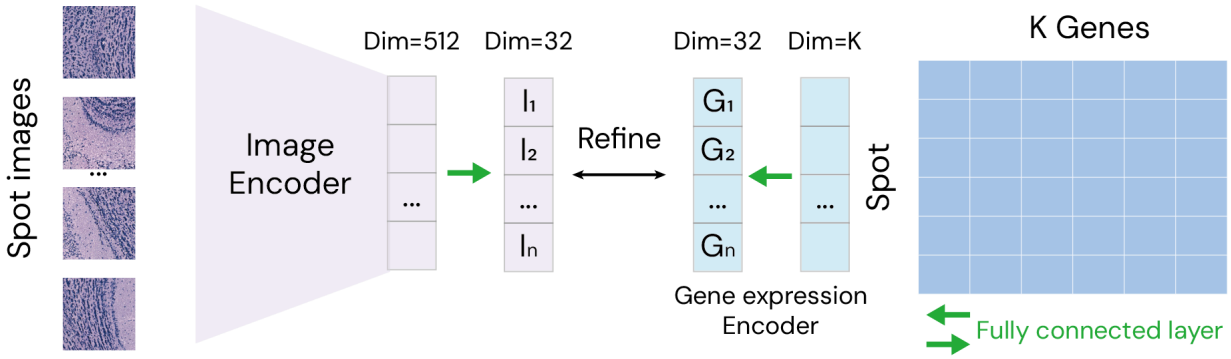

Figure R3 Fine-tune model structure. We replaced the text encoder in PLIP and CLIP with a single fully connected layer (gene expression encoder). We also add a fully connected layer after the image encoder to make the dimension same as the gene expression encoder, which in our case is 32.

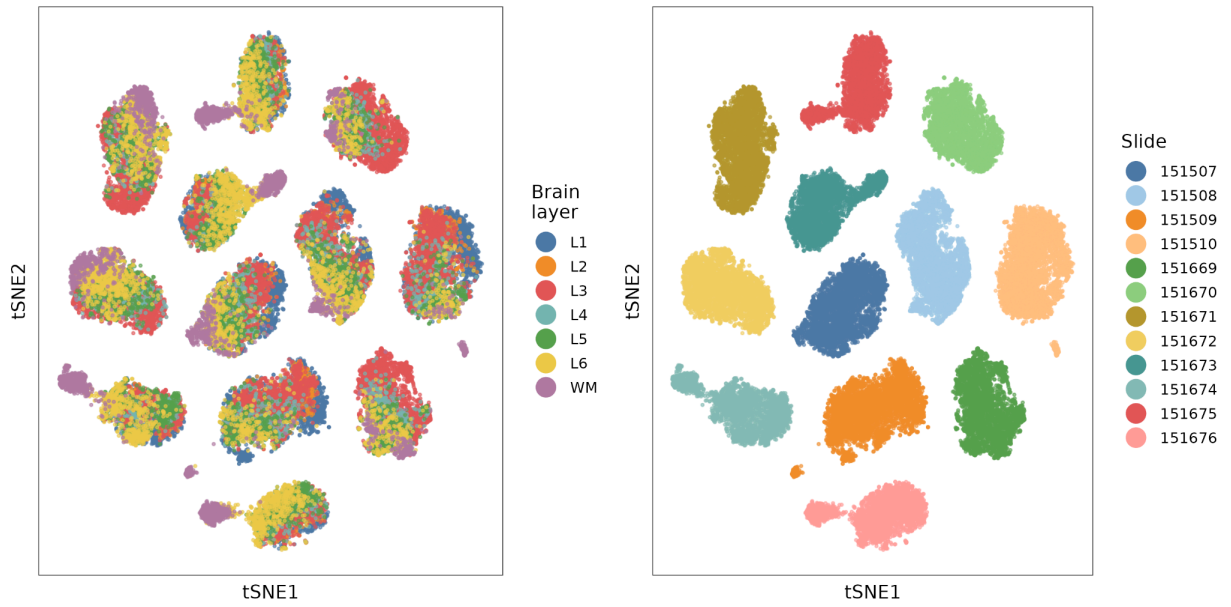

Figure R4 Batch effect in Maynard et al. human brain gene expression with hvg gene set. Left: gene expression tSNE embedding colored by brain layers. Right: gene expression tSNE embedding colored by slide name.

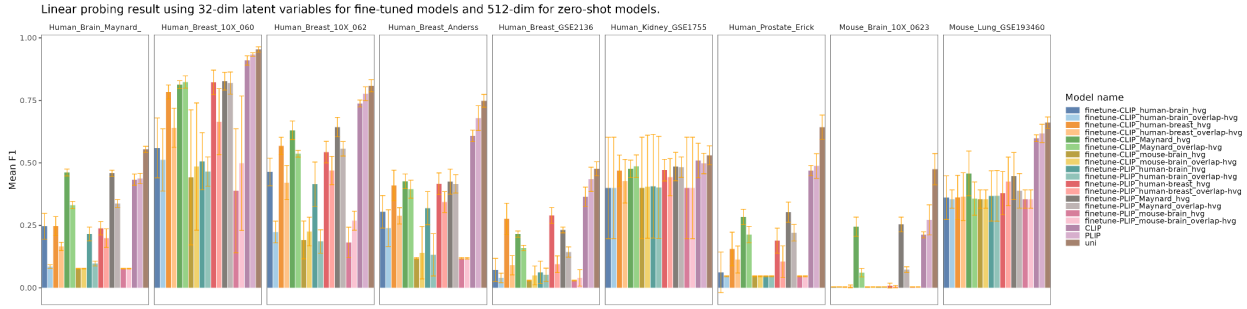

Figure R5 Classification results evaluated on multiple dataset for model fine-tuned with various training data. We fine-tuned the CLIP and PLIP models with human brain, human breast, Maynard et al., mouse brain. For fine-tuned models, we evaluate the 32-dimensional image embedding in a classification task. For zero-shot models, we evaluate the 512-dimensional image embedding. The classification performance of these models are tested on 9 datasets with annotation. The bars are colored by model type. The model name of the fine-tuned models is in “fine-tune model-type training-datatype gene-set” format.

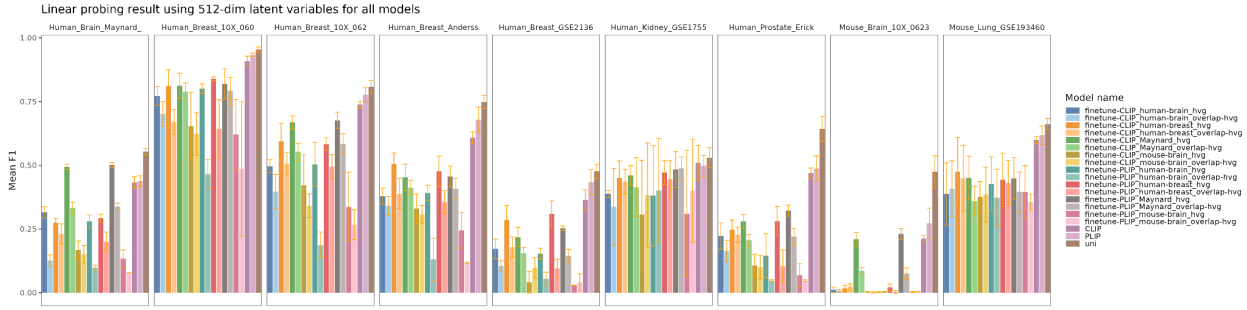

Figure R6 Classification results evaluated on multiple dataset for model fine-tuned with various training data. We fine-tuned the CLIP and PLIP models with human brain, human breast, Maynard et al., mouse brain. For both fine-tuned models and zero-shot models, we evaluate the 512-dimensional image embedding. The classification performance of these models are tested on 9 datasets with annotation. The bars are colored by model type. The model name of the fine-tuned models is in “fine-tune model-type training-datatype gene-set” format.