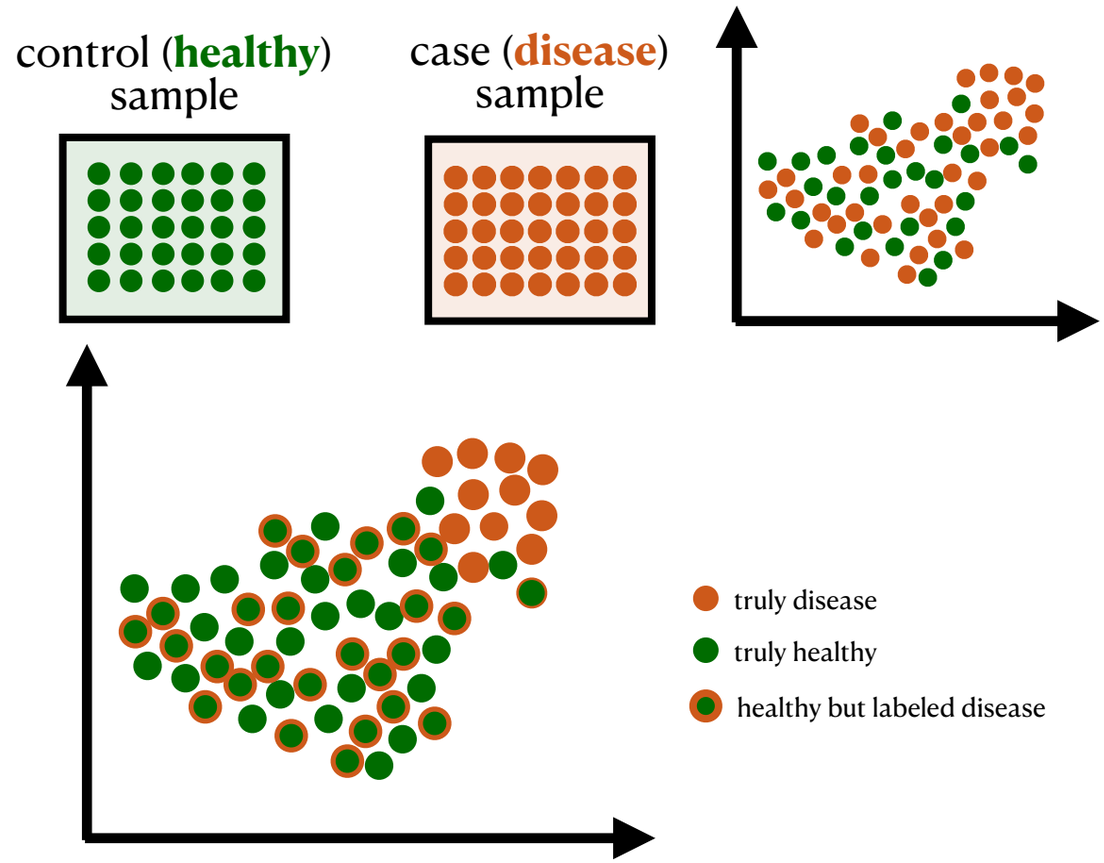
Label Refinement via Representation Augmentation for Boosting Disease Marker Genes Identification

ALEKSANDRINA GOEVA[1]
JONAH LANGLIEB[1]
EVAN MACOSKO[1, 2]

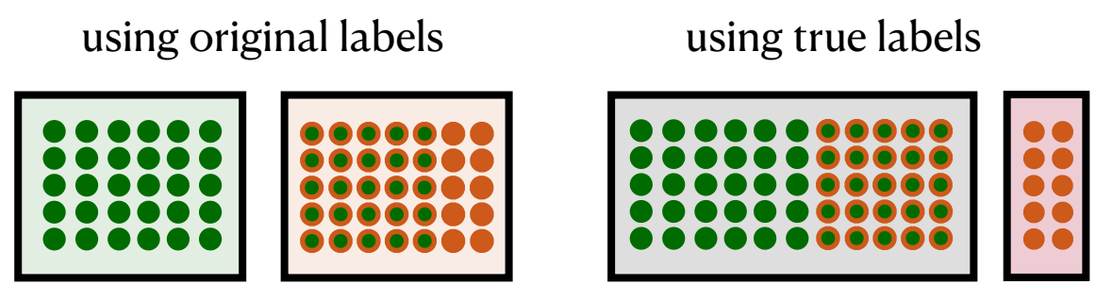
[1] Broad Institute of Harvard and MIT
[2] Department of Psychiatry, MGH

BROAD INSTITUTE
SIXTEENTH ANNUAL RETREAT
DECEMBER 14 – 17, 2020

What if only some of the cell in a disease sample are affected by the disease?

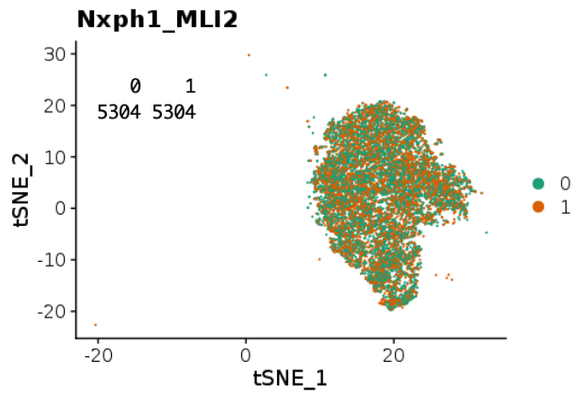


Differential Gene Expression

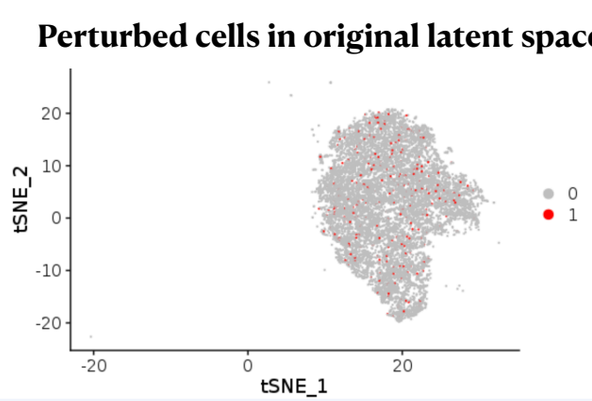


Disease signatures hiding in plain sight

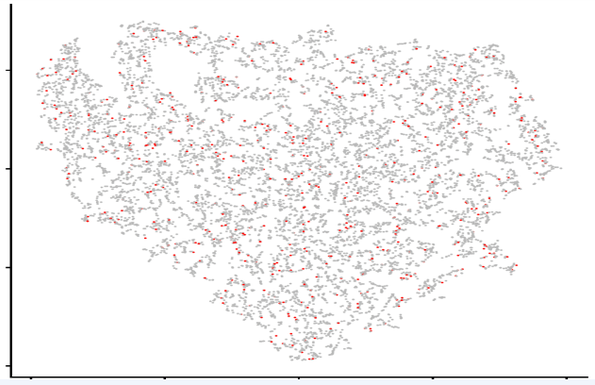
BROAD INSTITUTE
SIXTEENTH ANNUAL RETREAT
DECEMBER 14 – 17, 2020



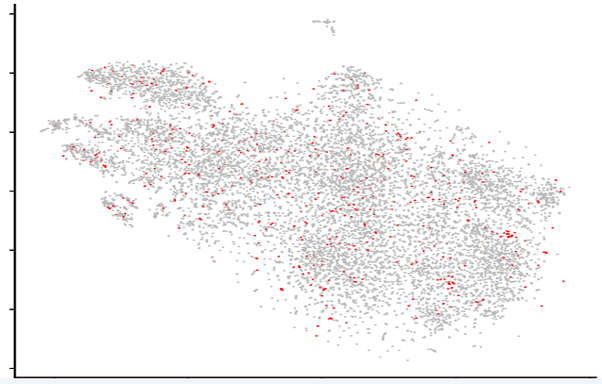
randomly split a **homogeneous** cell type into case (**disease**) and control (**healthy**)



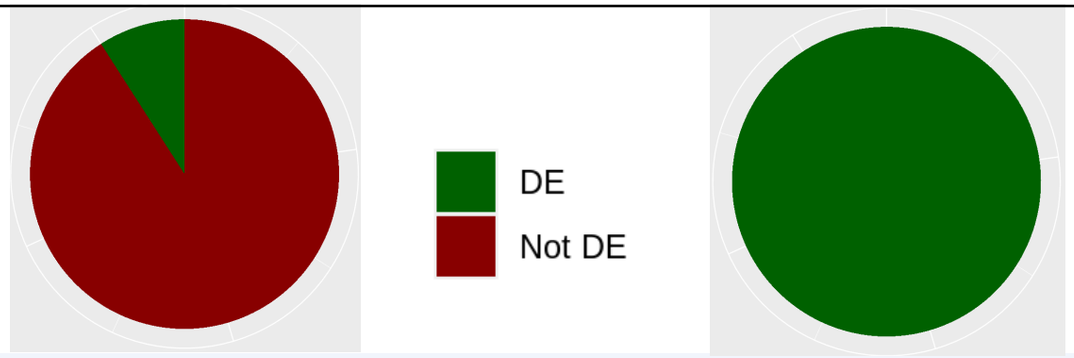
select a **fraction of the disease**-labeled cells to be perturbed (**5%**)



despite large magnitude of the perturbation, the latent space **does not separate**



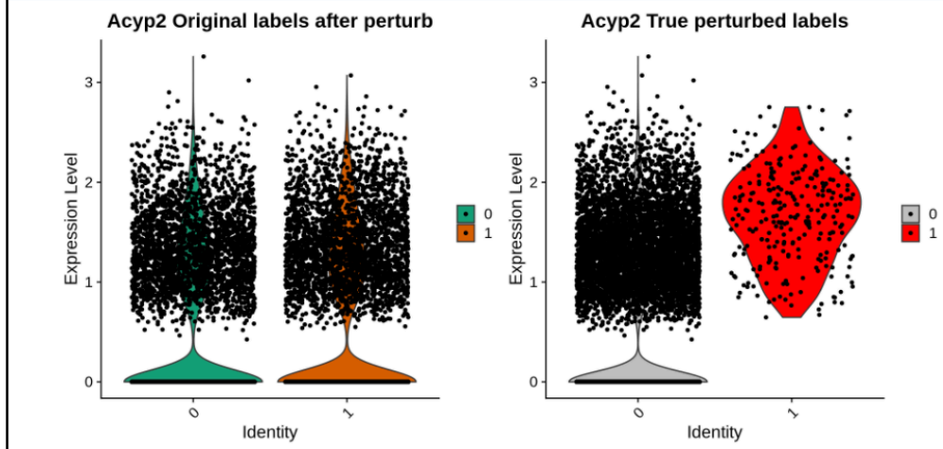
even if we include **more PCs** (tSNE based on top 50 PCs)

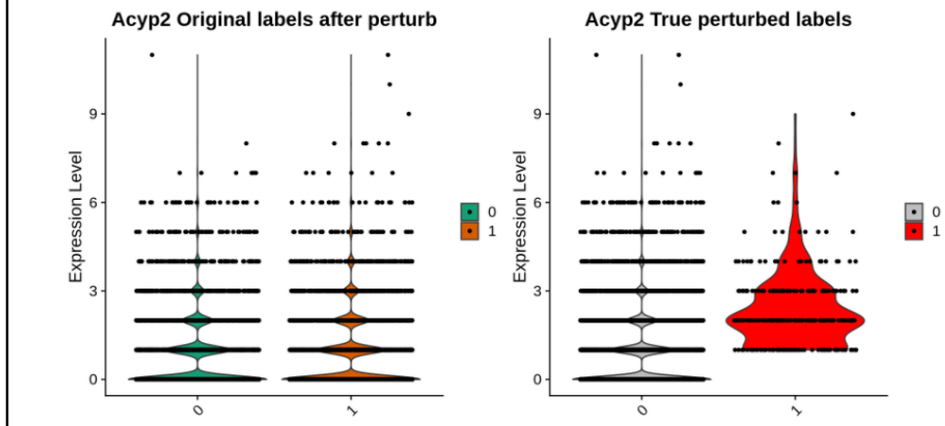


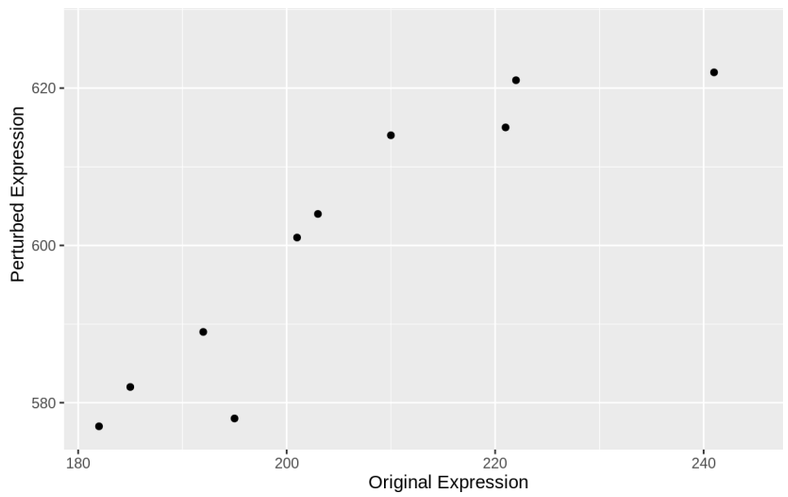
DE using case/control fails to find the perturbed genes

but it **succeeds given the true perturbation labels**

magnitude of the perturbation



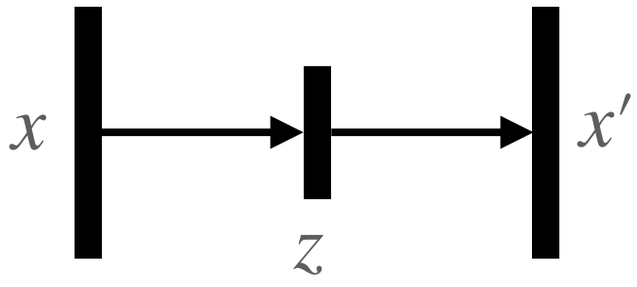




How can we correct the disease labels to reflect the true perturbation?

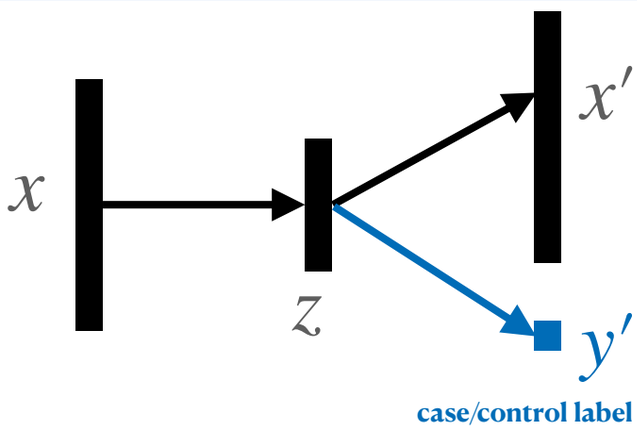
BROAD INSTITUTE
SIXTEENTH ANNUAL RETREAT
DECEMBER 14 – 17, 2020

standard representation learning



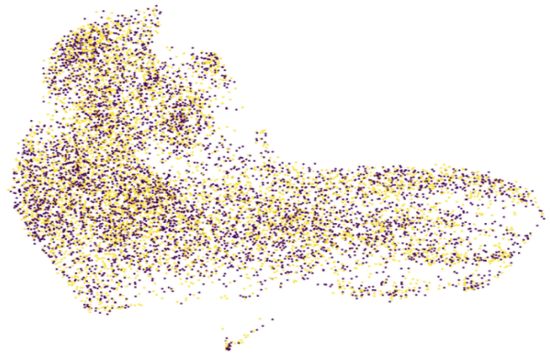
$$\text{loss} = \text{MSE}(x, x')$$

shaping the latent space using metadata



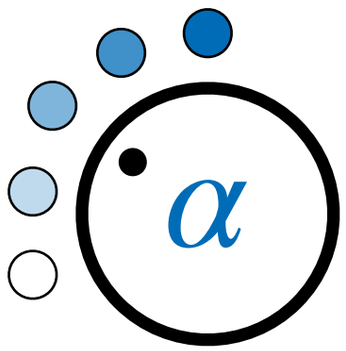
$$\text{loss} = \text{MSE}(x, x') + \alpha\text{BCE}(y, y')$$

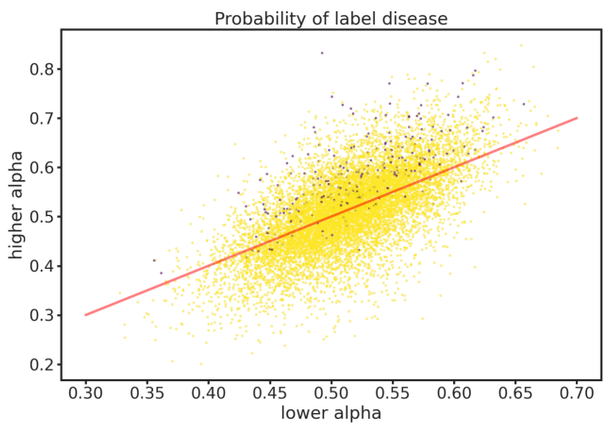
case/control labels



true perturbation labels

highlighting the truly perturbed cells





We can compute a predicted probability of disease label based on a logistic regression trained in the new latent space. Comparing the predicted probability as we increase the alpha parameter shows promise for highlighting the truly perturbed cells.
This is WIP, stay tuned for updates :)