Deep generative modelling of bulk transcriptomics

Iñigo Prada Luengo

Agenda

- Why are generative models useful?
  - The weather for my grandma & a scientist
  - Why are generative models useful in biology
- A generative model of the bulk human transcriptome
  - Why even model bulk RNA-seq data?
  - A generative model for bulk RNA-seq
  - Detecting tumor origins
  - A new way of finding differentially expressed genes

Discriminative vs Generative models

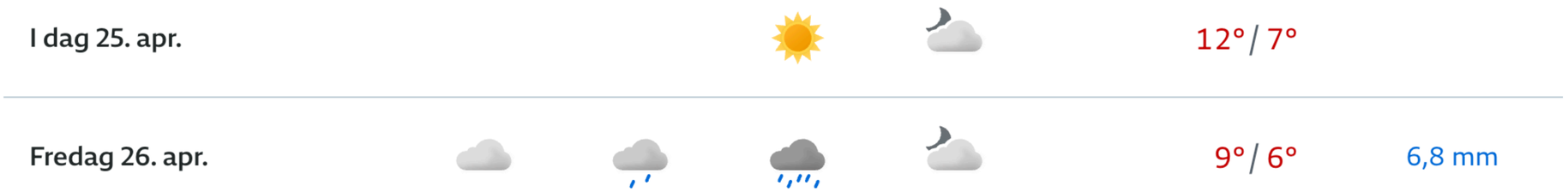
- A discriminative model: $P(y|x)$ "Lets learn to predict an outcome given the observations"
- A generative model: $P(y, x)$ "Lets learn a joint distribution of all variables"
- Why are generative models attractive?
- We learn a model of how our data is *generated*.

Modeling the weather

- A discriminative model for the weather: $P(y|x)$

Modeling the weather

- A discriminative model for the weather: $P(y|x)$



Modeling the weather

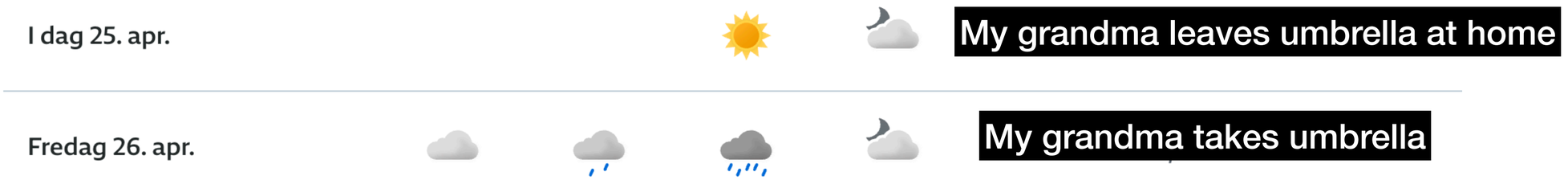
- A discriminative model for the weather: $P(y|x)$

I dag 25. apr.  My grandma leaves umbrella at home

Fredag 26. apr.  My grandma takes umbrella

Modeling the weather

- A discriminative model for the weather: $P(y|x)$



- A generative model for the weather: $P(y, x)$

Temperature

Humidity

Wind

⋮

Precipitation

Sunshine

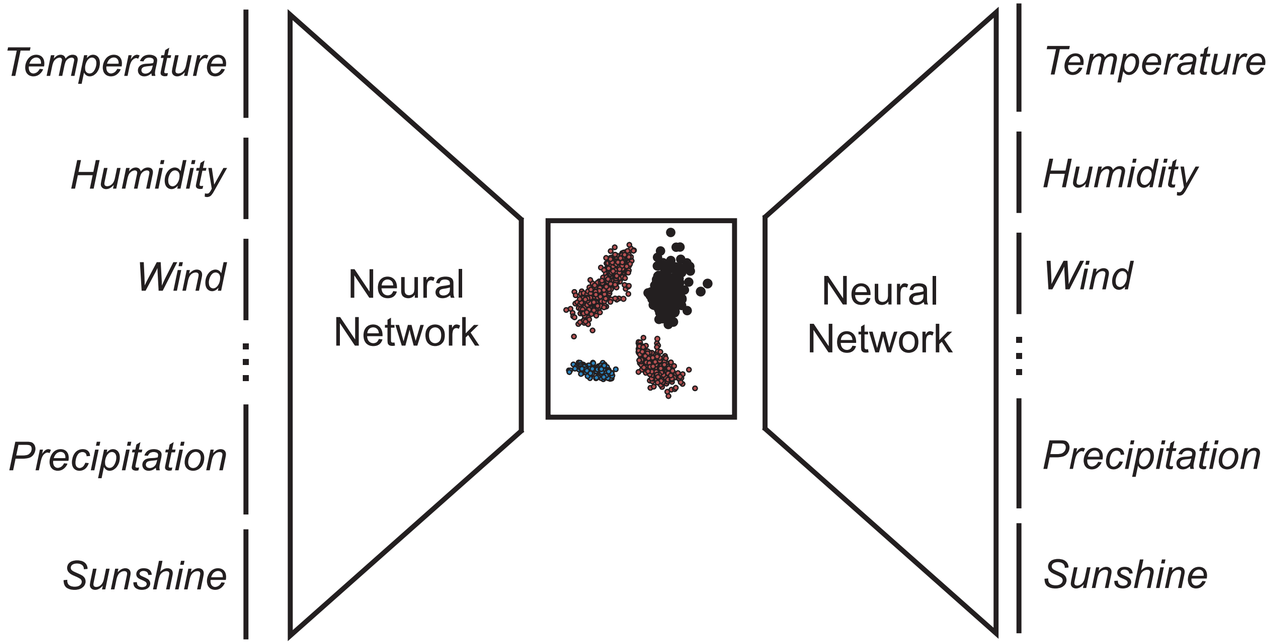
Modeling the weather

- A discriminative model for the weather: $P(y|x)$

I dag 25. apr. ☀️ 🌥️ My grandma leaves umbrella at home

Fredag 26. apr. ☁️ 🌧️ 🌧️ 🌥️ My grandma takes umbrella

- A generative model for the weather: $P(y, x)$



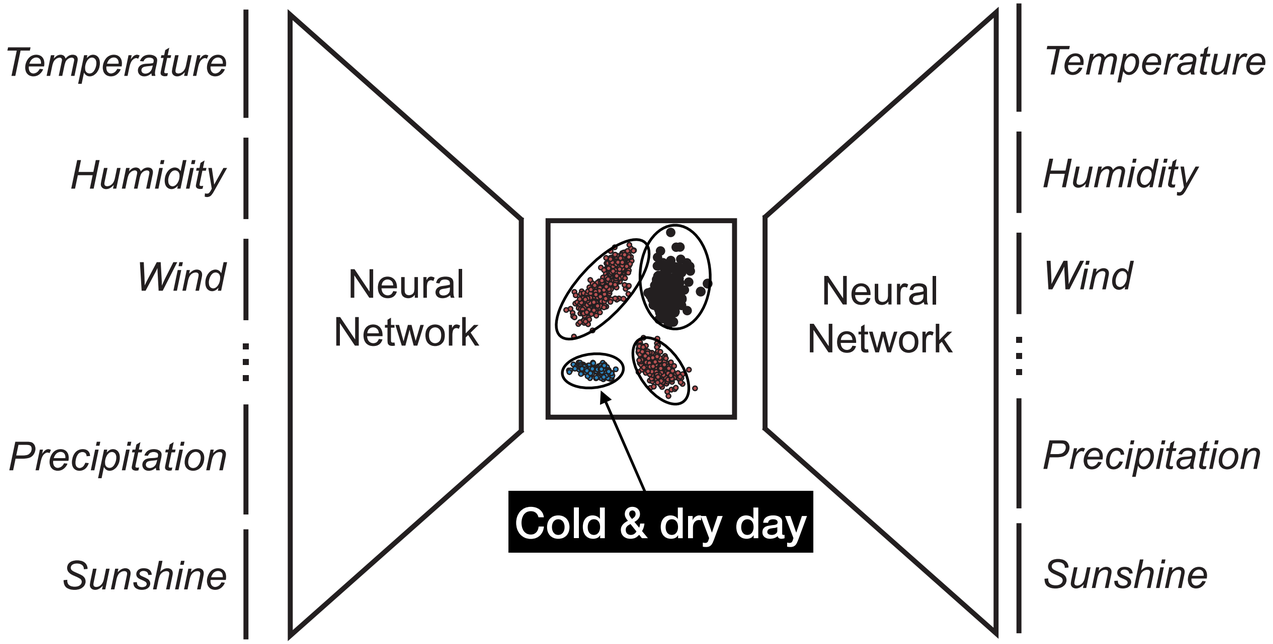
Modeling the weather

- A discriminative model for the weather: $P(y|x)$

I dag 25. apr. ☀️ 🌥️ My grandma leaves umbrella at home

Fredag 26. apr. ☁️ 🌧️ 🌧️ 🌥️ My grandma takes umbrella

- A generative model for the weather: $P(y, x)$



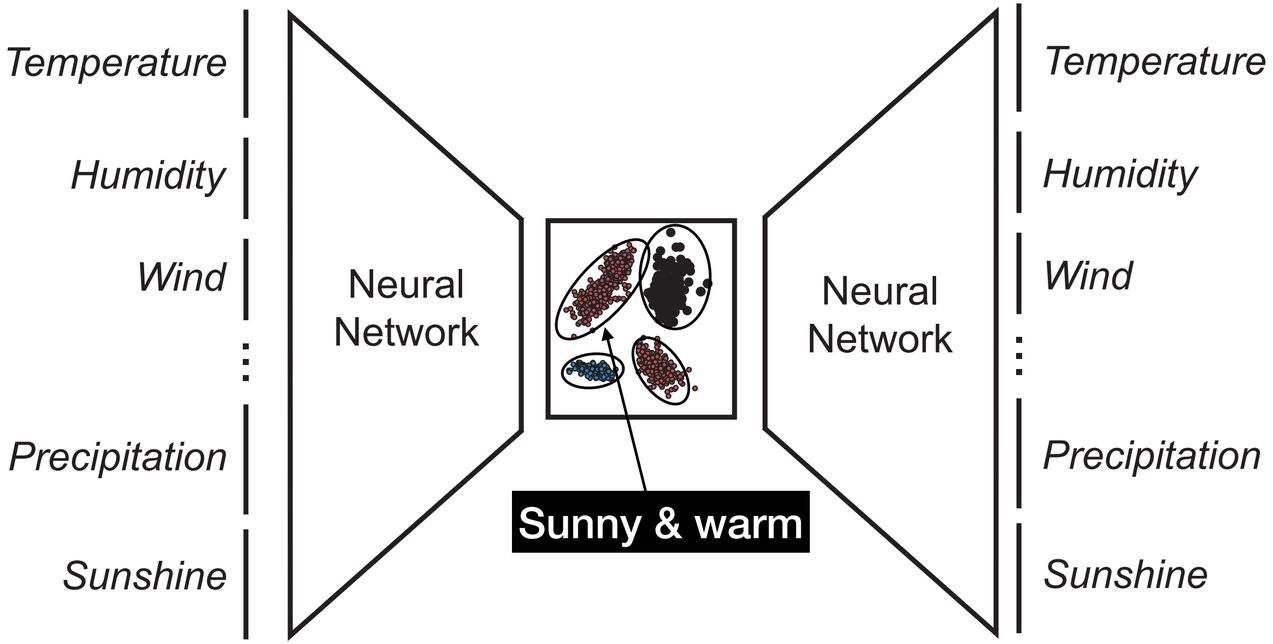
Modeling the weather

- A discriminative model for the weather: $P(y|x)$

I dag 25. apr. ☀️ 🌥️ My grandma leaves umbrella at home

Fredag 26. apr. ☁️ 🌧️ 🌧️ 🌥️ My grandma takes umbrella

- A generative model for the weather: $P(y, x)$



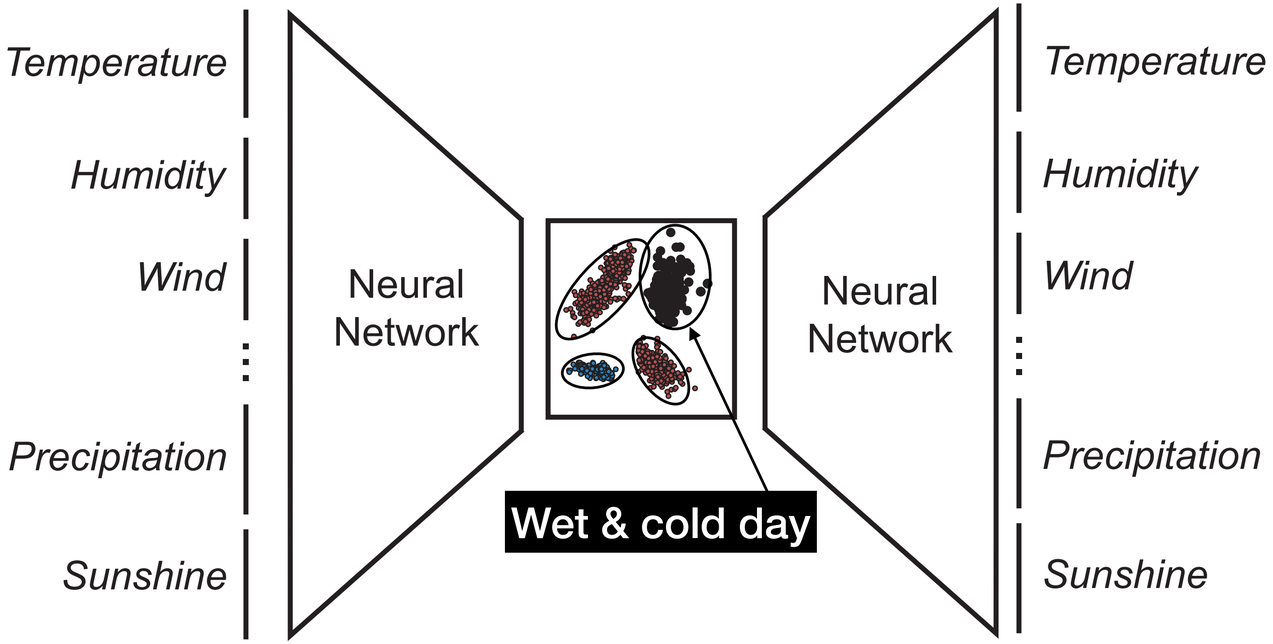
Modeling the weather

- A discriminative model for the weather: $P(y|x)$

I dag 25. apr. ☀️ 🌥️ My grandma leaves umbrella at home

Fredag 26. apr. ☁️ 🌧️ 🌧️ 🌥️ My grandma takes umbrella

- A generative model for the weather: $P(y, x)$



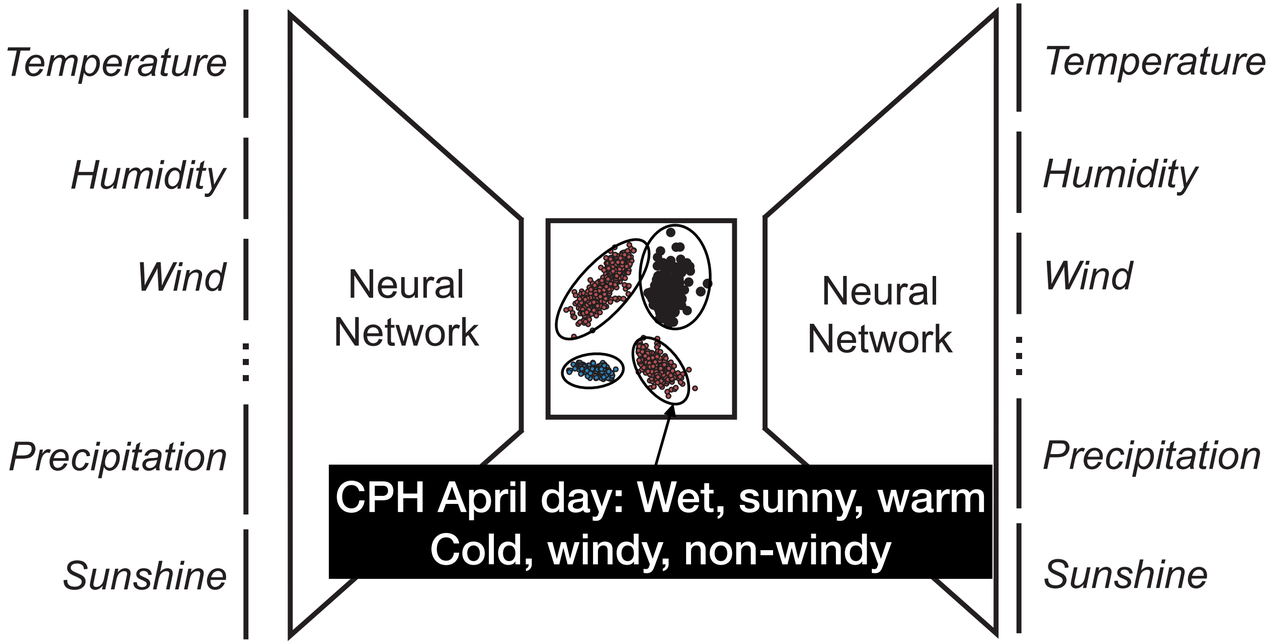
Modeling the weather

- A discriminative model for the weather: $P(y|x)$

I dag 25. apr. ☀️ 🌥️ My grandma leaves umbrella at home

Fredag 26. apr. ☁️ 🌧️ 🌧️ 🌥️ My grandma takes umbrella

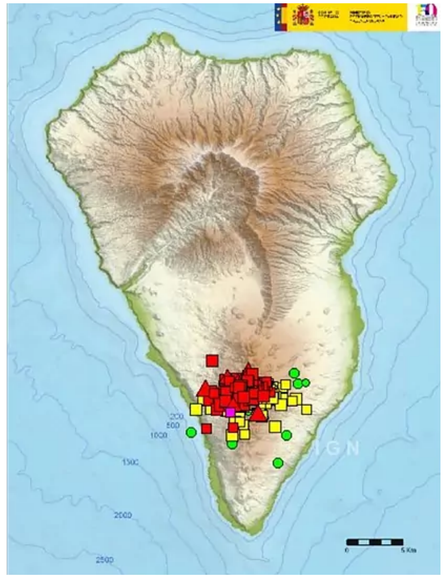
- A generative model for the weather: $P(y, x)$



Generative models for science

- We can use generative models to understand the world

2021 La Palma eruption, Canary Islands. Spain



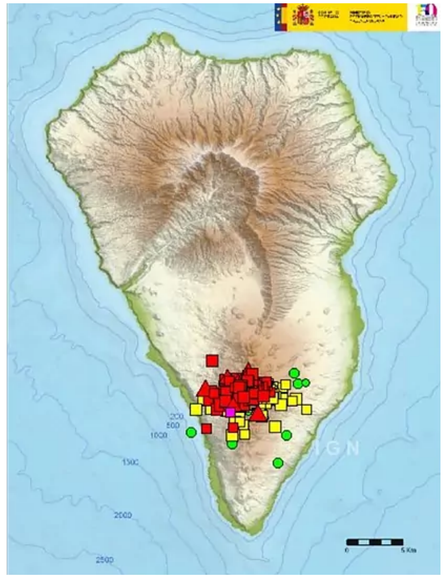
Source: Spanish Geology Institute

Disclaimer: I am neither a geologist nor a meteorologist. These are simply illustrative examples

Generative models for science

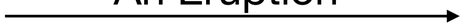
- We can use generative models to understand the world

2021 La Palma eruption, Canary Islands. Spain



Source: Spanish Geology Institute

Generative Model
Of
An Eruption



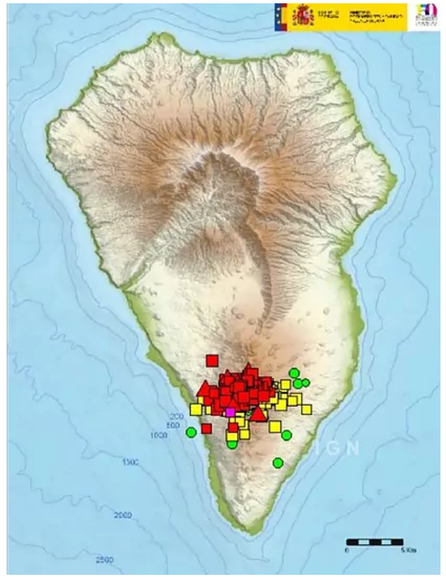
Use knowledge elsewhere

Disclaimer: I am neither a geologist nor a meteorologist. These are simply illustrative examples

Generative models for science

- We can use generative models to understand the world

2021 La Palma eruption, Canary Islands. Spain

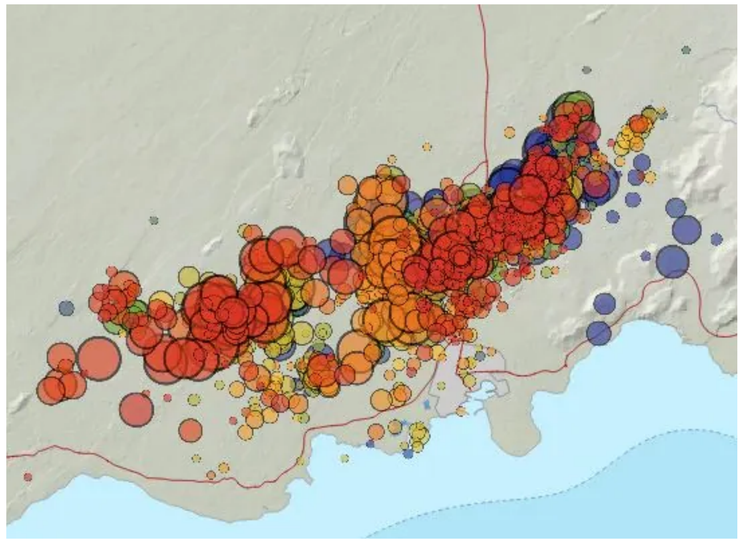


Source: Spanish Geology Institute

Generative Model
Of
An Eruption

Use knowledge elsewhere

2024 Grindavik eruption, Iceland



Source: Icelandic Meteo Office

Disclaimer: I am neither a geologist nor a meteorologist. These are simply illustrative examples

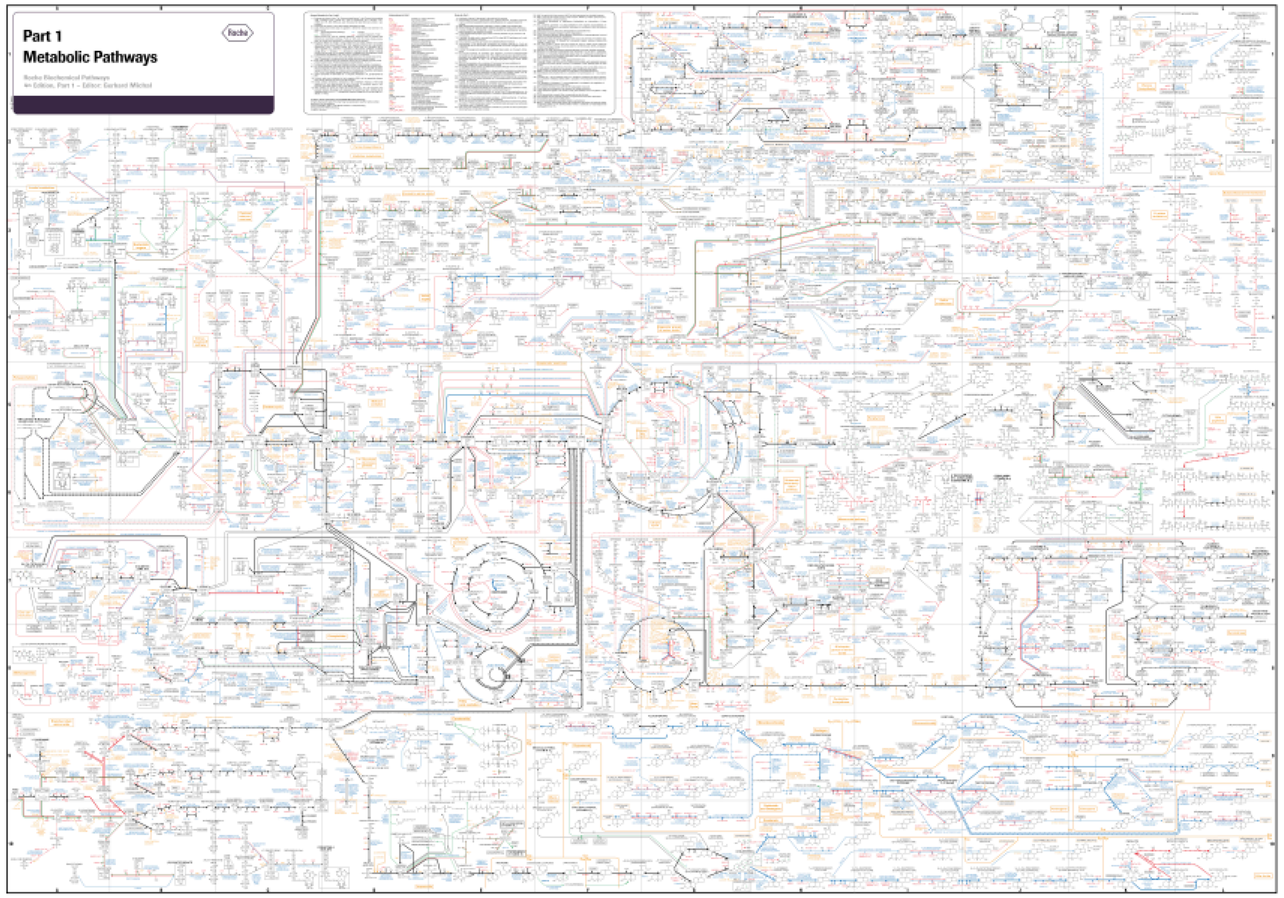
Why are generative models useful for biology?

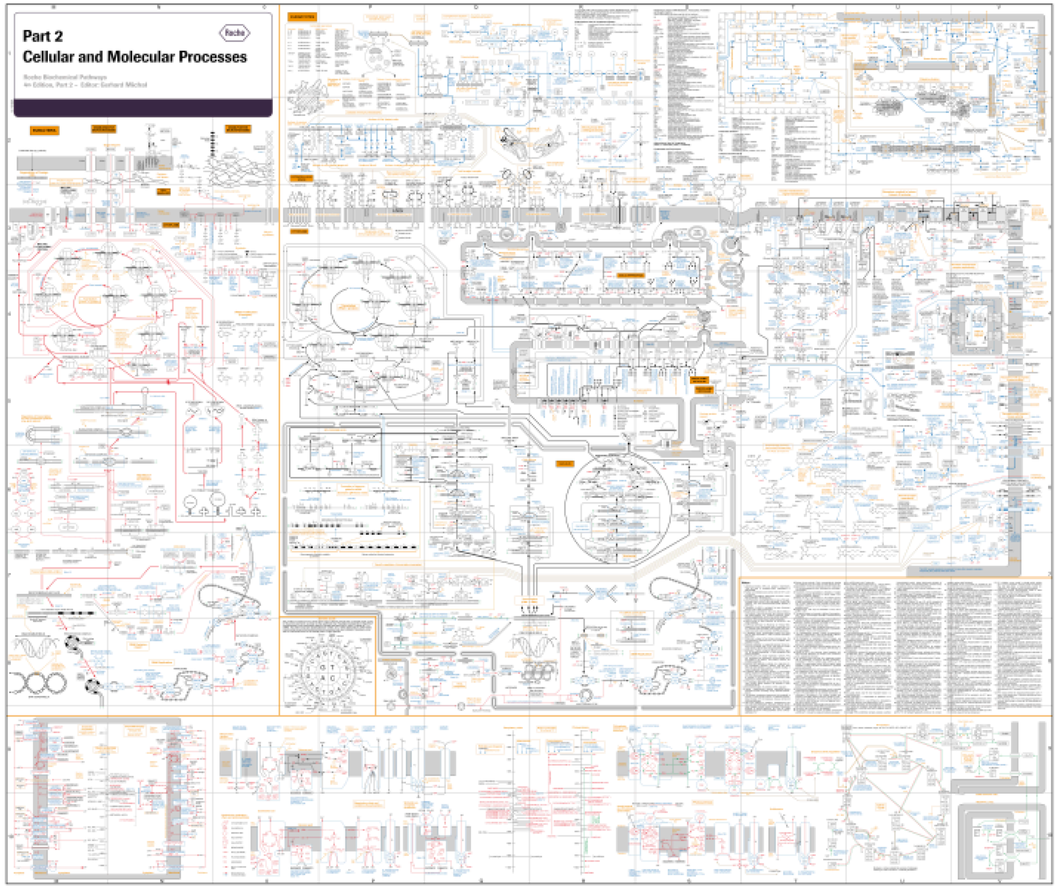
Why are generative models useful for biology?

- Molecular biology is high dimensional & interconnected. How do we make sense of it?
- Idea:let’s try to make an schematic, extremely oversimplified model of what is going on in a cell

Why are generative models useful for biology?

- Molecular biology is high dimensional & interconnected. How do we make sense of it?
- Idea:let's try to make an schematic, extremely oversimplified model of what is going on in a cell

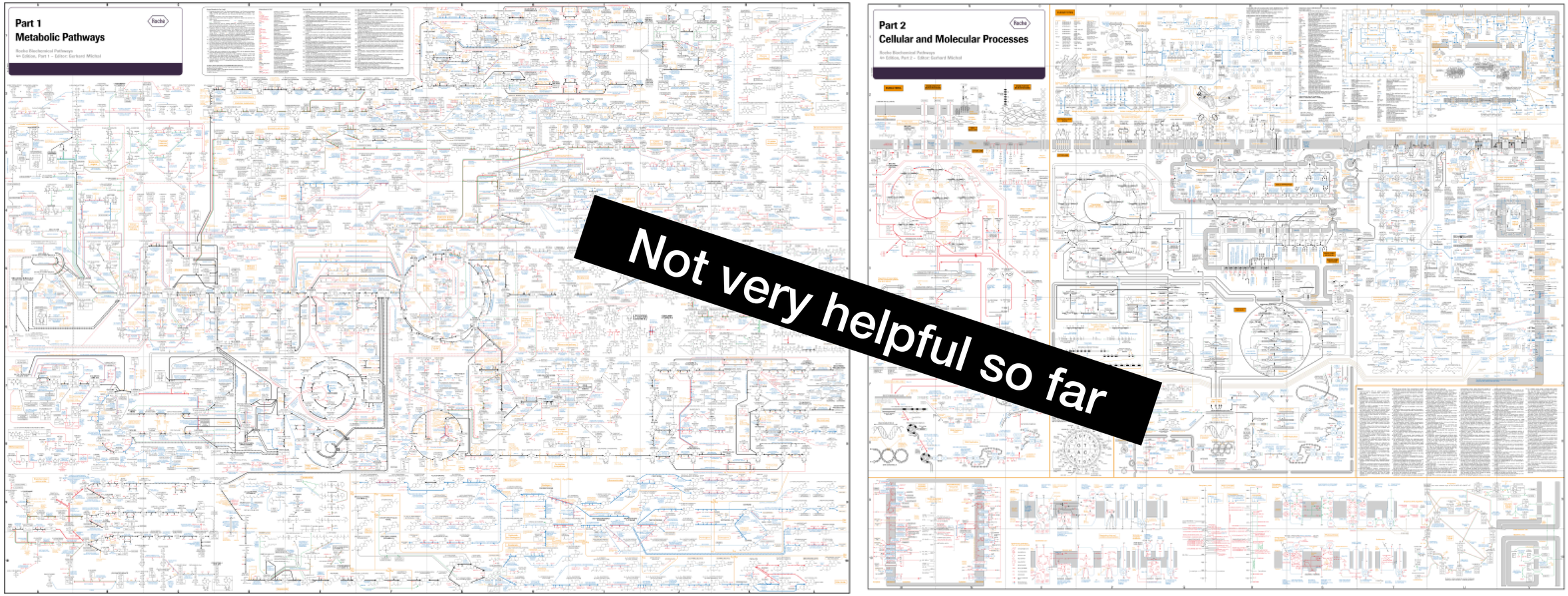




Source: Roche

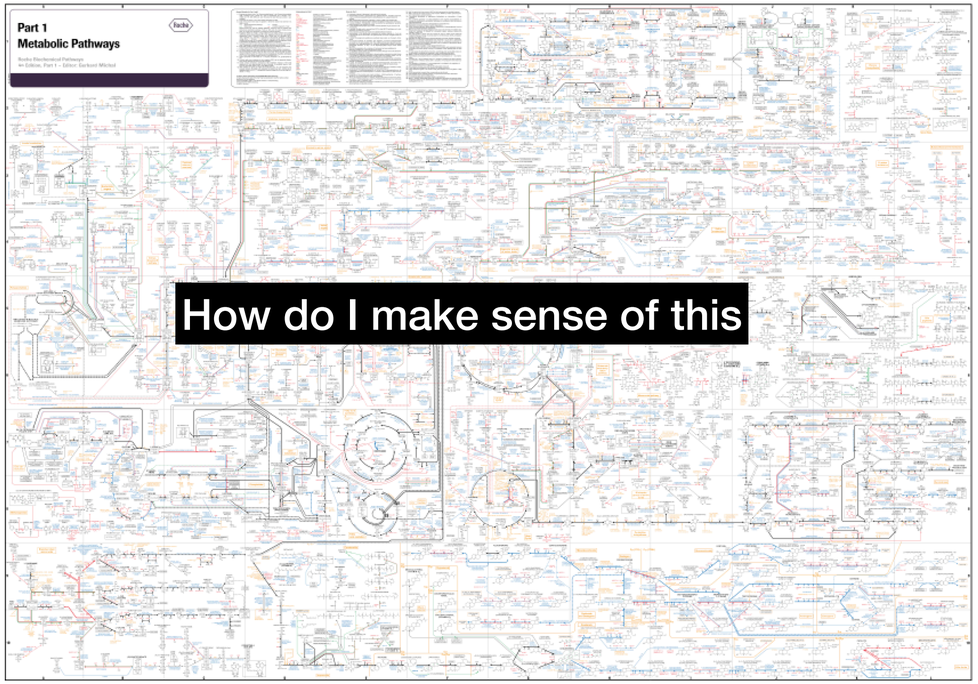
Deep generative models for biology

- Molecular biology is high dimensional & interconnected. How do we make sense of it?
- Idea:let's try to make an schematic, extremely oversimplified model of what is going on in a cell



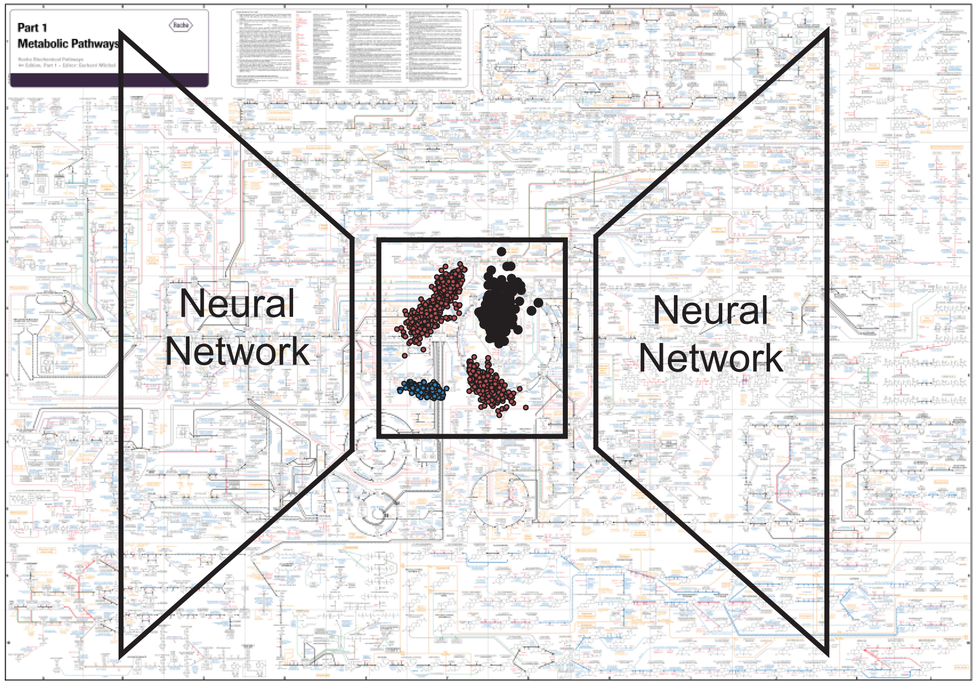
Source: Roche

Why don’t we let a computer figure this out?



- We nowadays can collect vast amounts of data using omic techniques

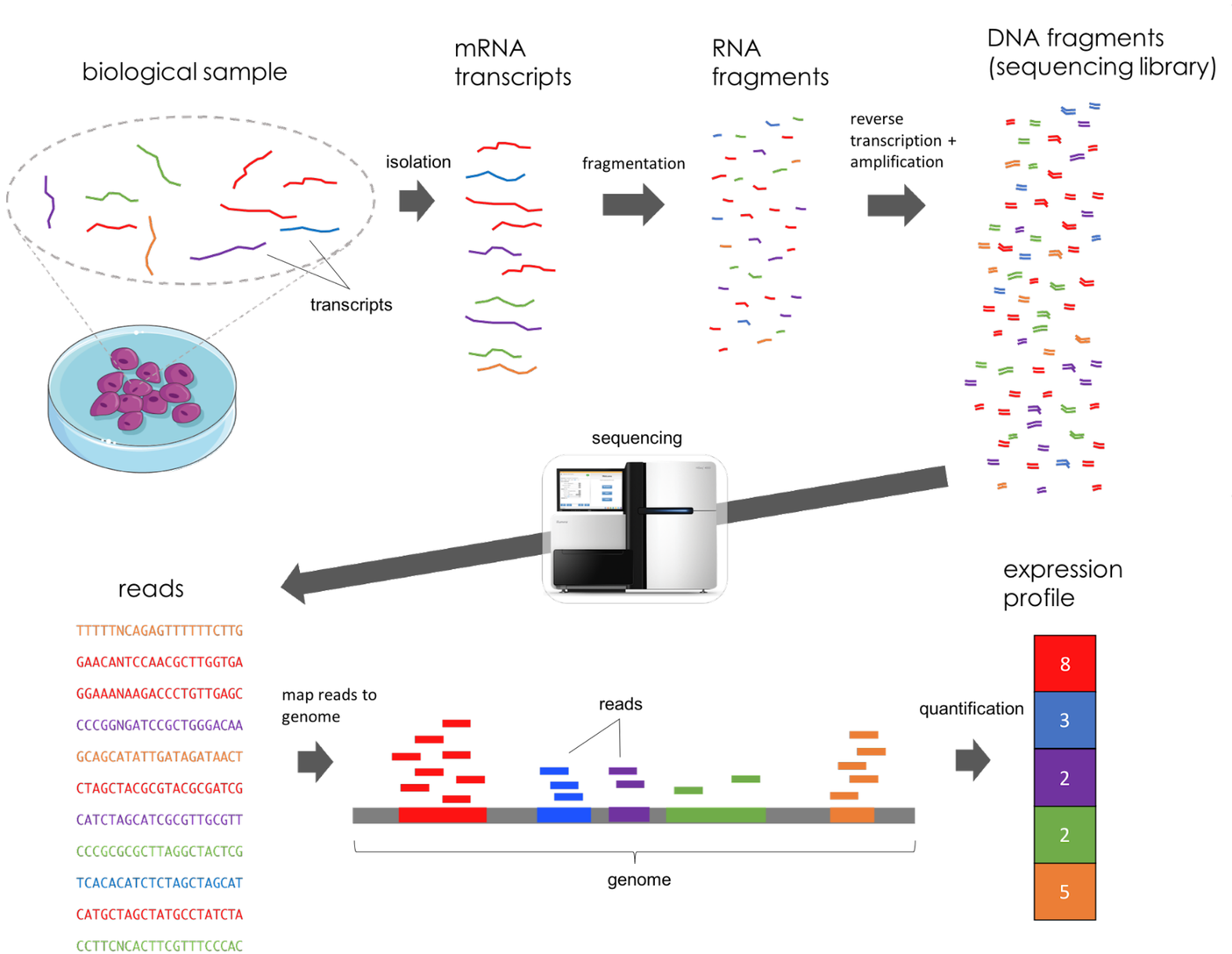
Why don’t we let a computer figure this out?



- We nowadays can collect vast amounts of data using omic techniques
- We can leave the task of figuring out all this associations to a neural network
- Highly successful in recent years
  - E.g. AlphaFold, scRNA-seq

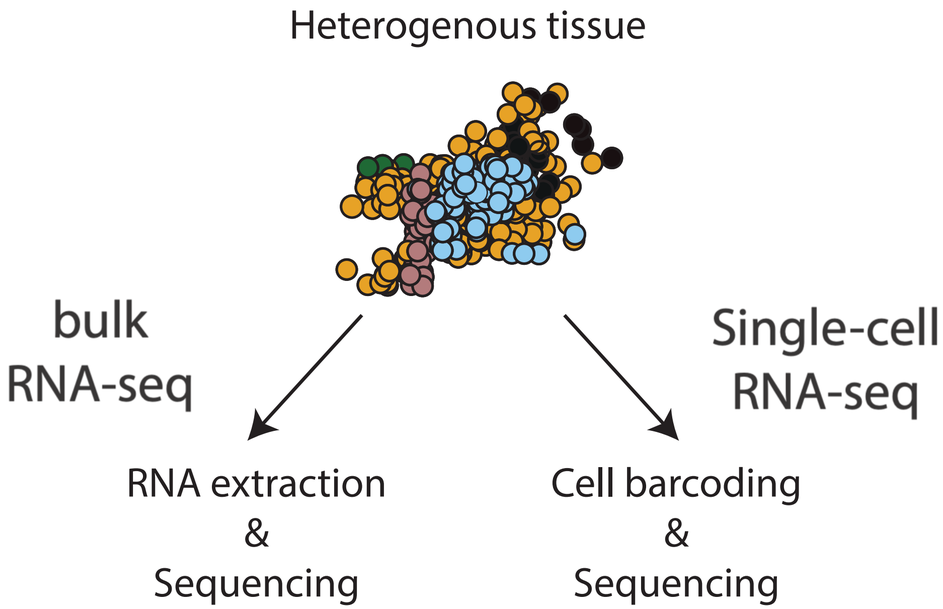
Deep generative modelling of bulk RNA-seq

The RNA-seq workflow

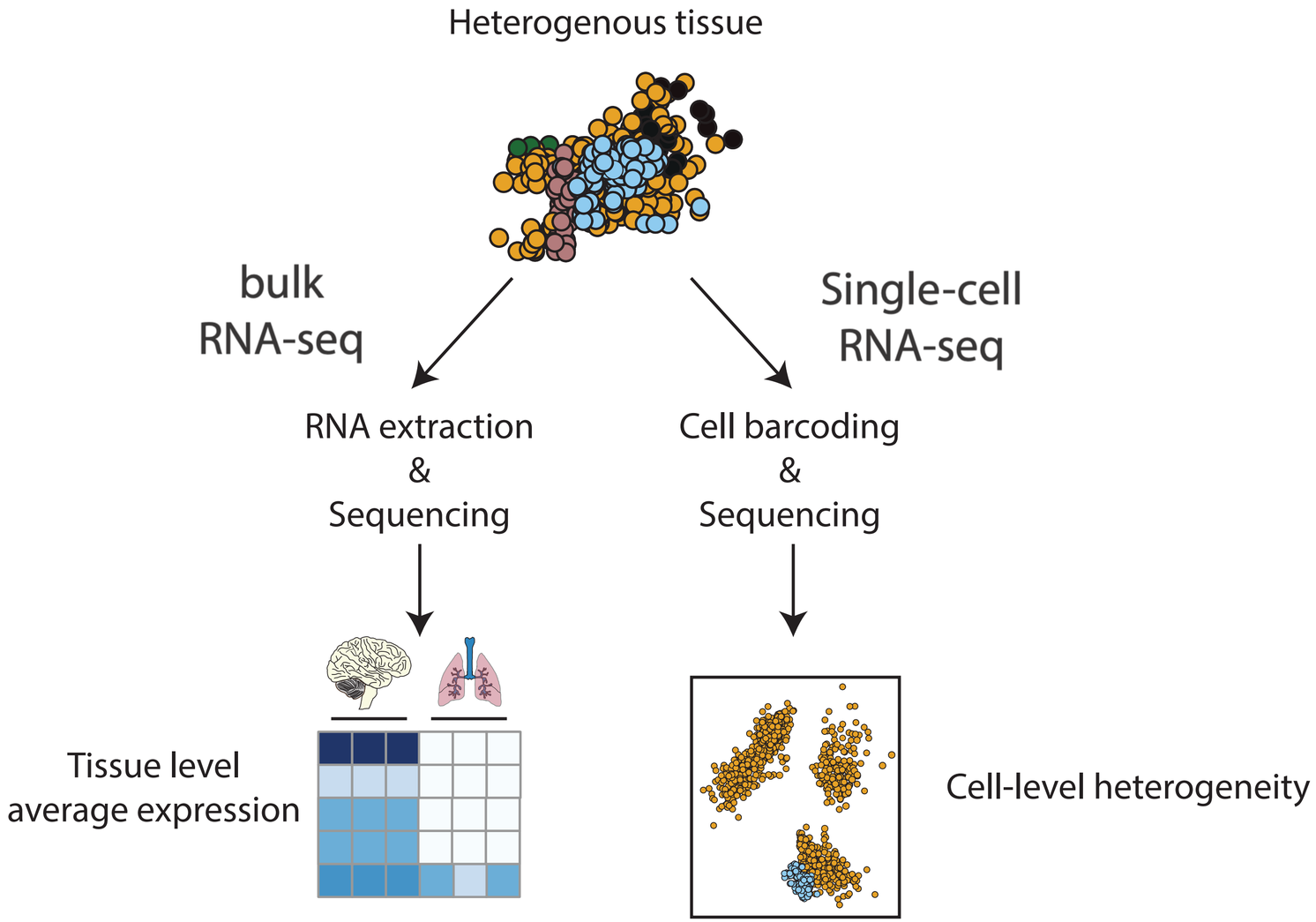


- Highly versatile technique:
  - Isoform, fusion gene detection…
- Main use is the detection of differentially expressed genes
  - E.g. Limma, DEseq2, EdgeR

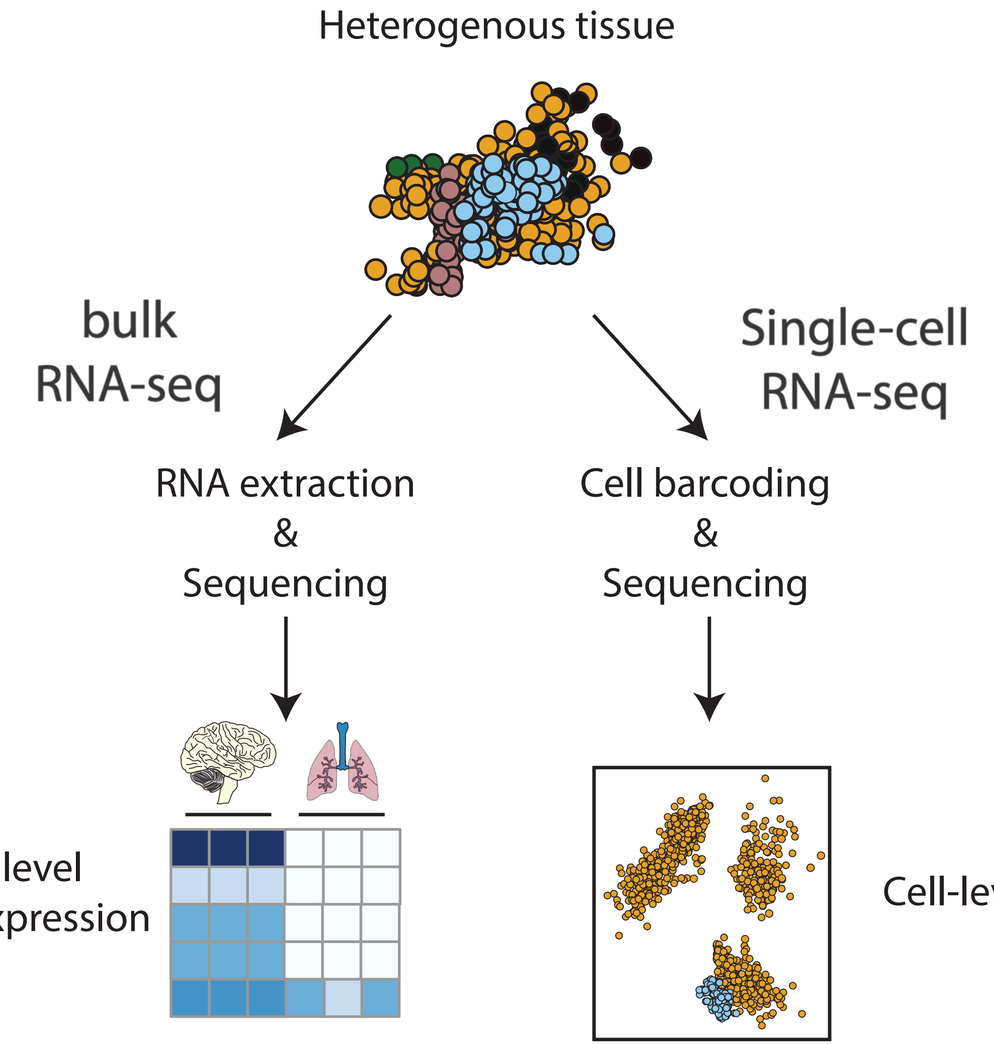
Why bulk RNA-seq?



Why bulk RNA-seq?

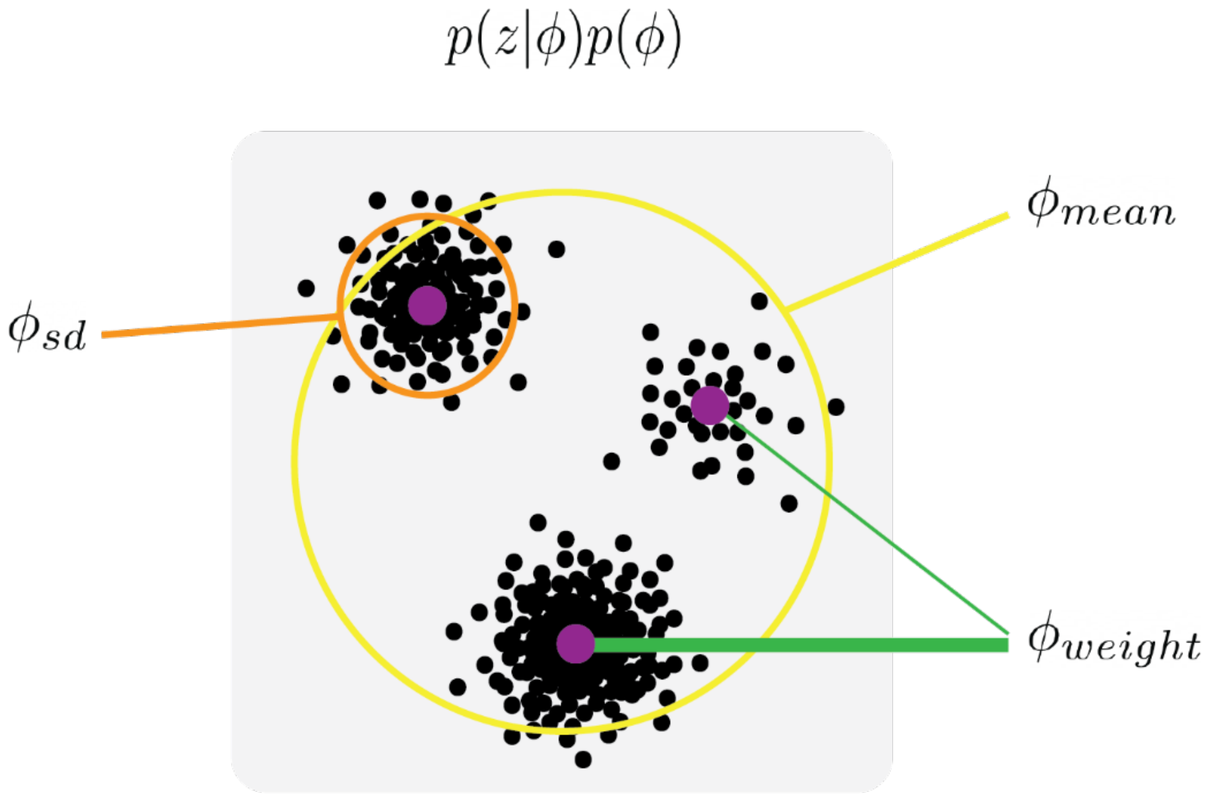


Why bulk RNA-seq?

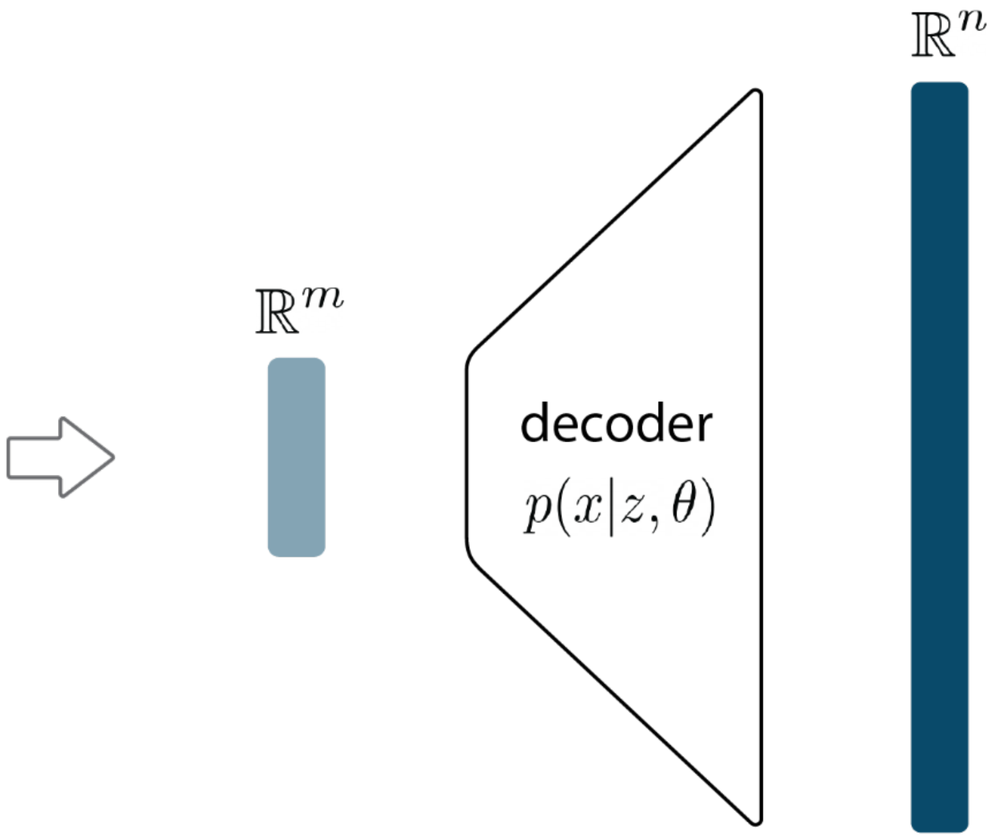


- Bulk RNA-seq
  - Mature technique
  - Less noisy
  - Main technique used in “production”

The Deep Generative Decoder



GMM

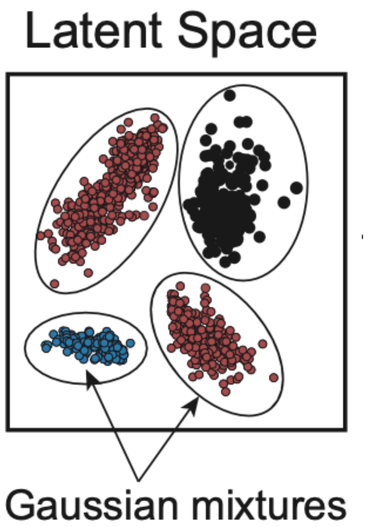


latent Z

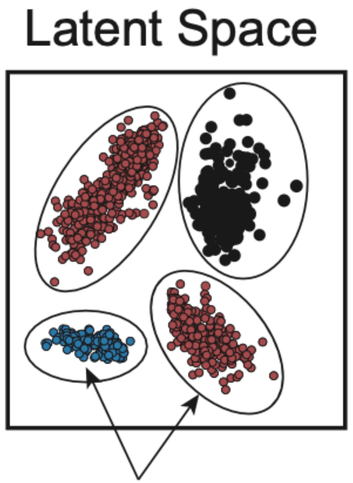
data X

Schuster & Krogh. Bioinformatics, 2023

Model set-up

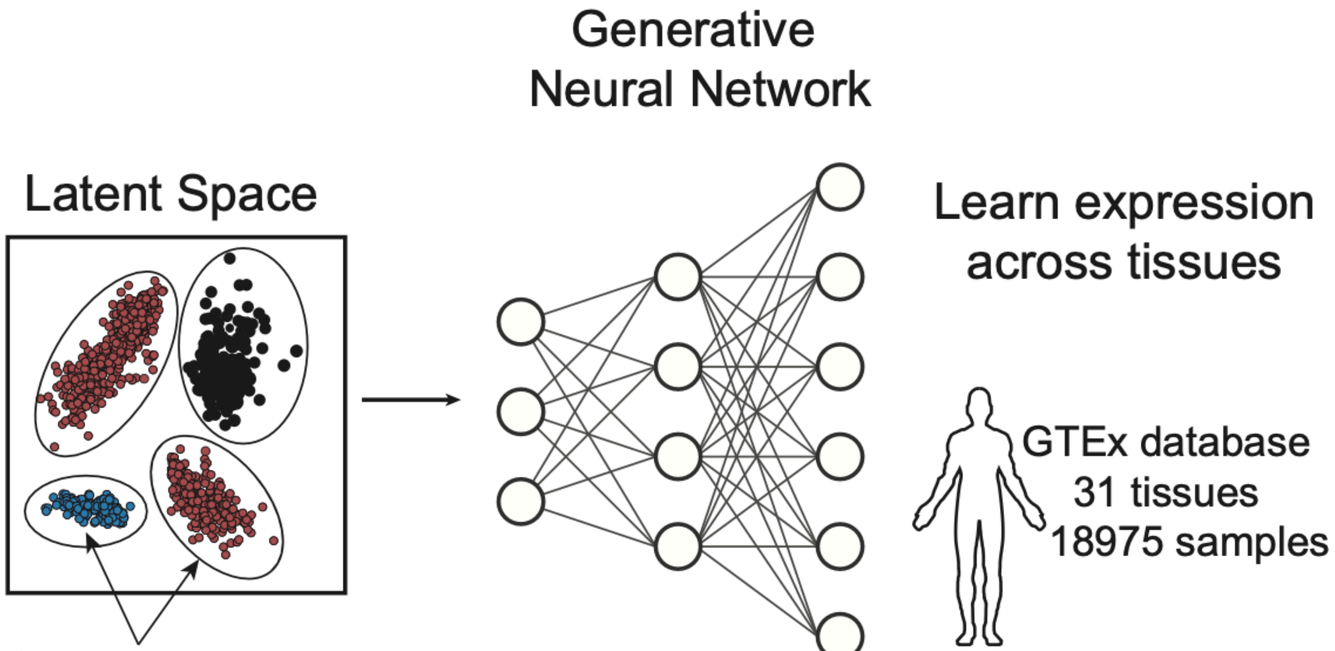


Model set-up



Latent space 50 dimensions
GMM with 45 components
Diagonal covariance

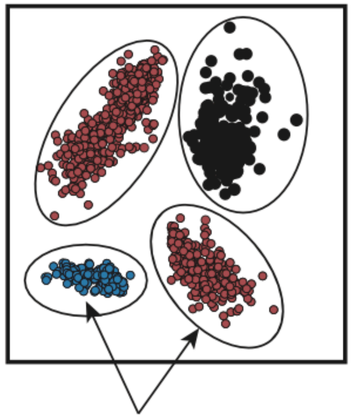
Model set-up


Generative
Neural Network
Latent Space
Learn expression
across tissues
GTEx database
31 tissues
18975 samples
Gaussian mixtures
Latent space 50 dimensions
GMM with 45 components
Diagonal covariance

Model set-up

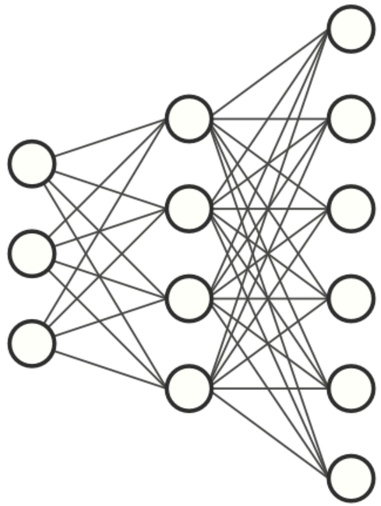
Generative
Neural Network

Latent Space



Gaussian mixtures

Latent space 50 dimensions
GMM with 45 components
Diagonal covariance

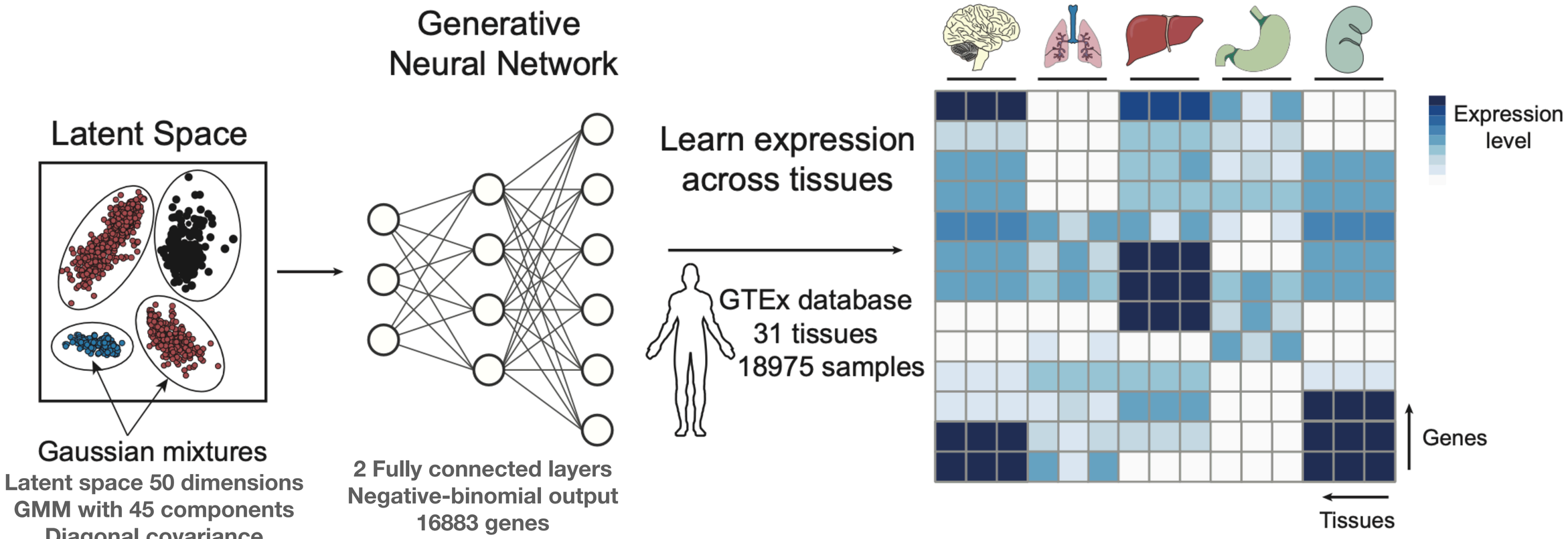


2 Fully connected layers
Negative-binomial output
16883 genes

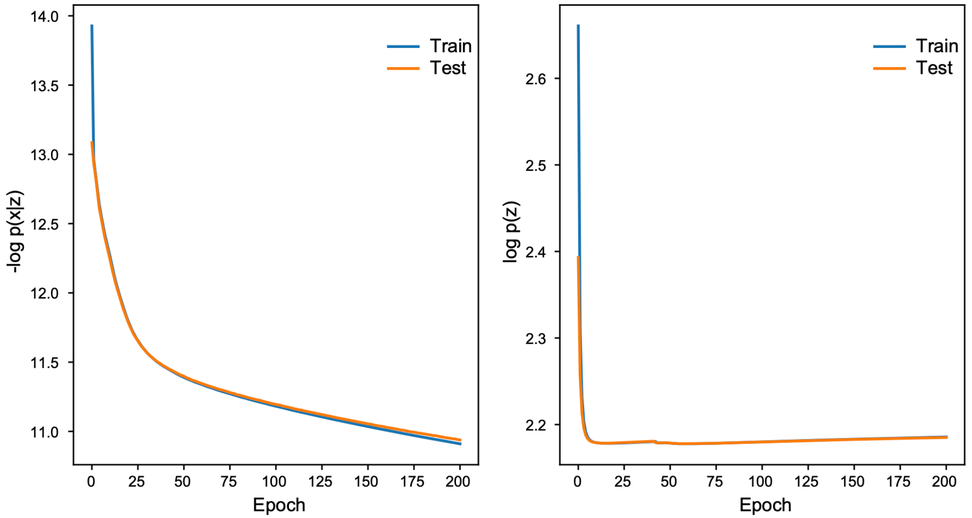
Learn expression
across tissues

GTEx database
31 tissues
18975 samples

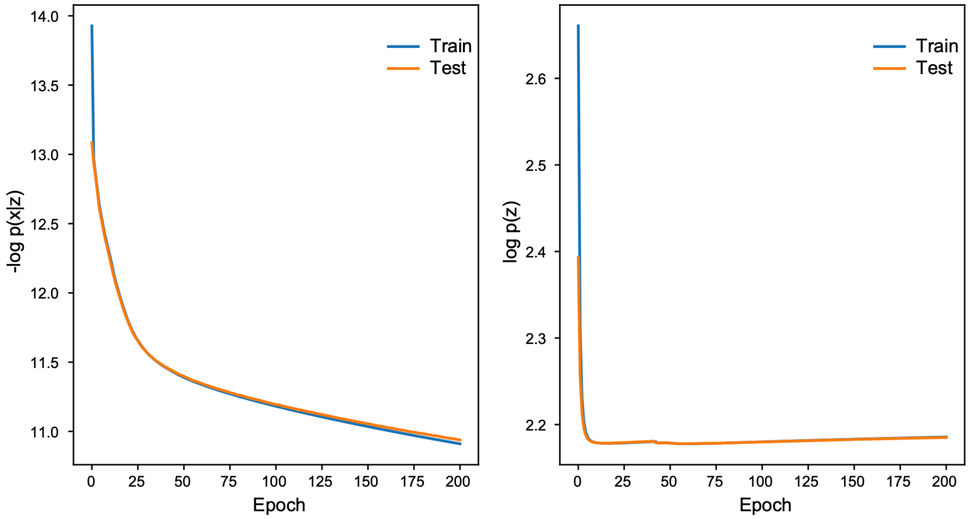
Model set-up

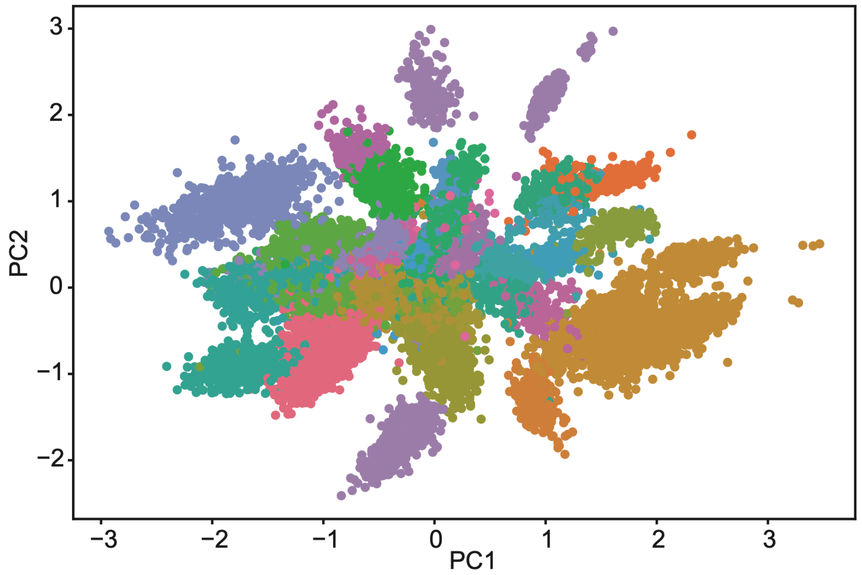

Latent Space
Gaussian mixtures
Latent space 50 dimensions
GMM with 45 components
Diagonal covariance
Generative
Neural Network
2 Fully connected layers
Negative-binomial output
16883 genes
Learn expression
across tissues
GTEx database
31 tissues
18975 samples
Expression
level
Genes
Tissues

Does the model learn?

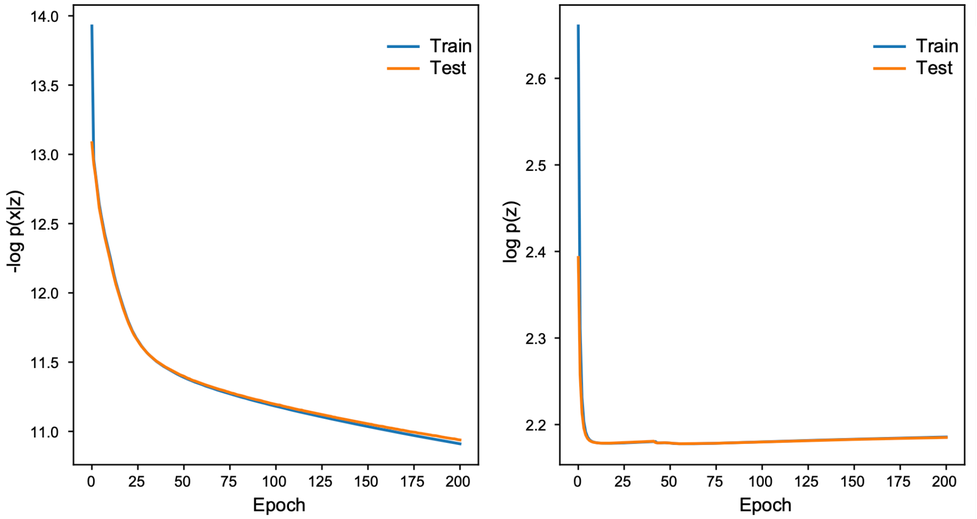


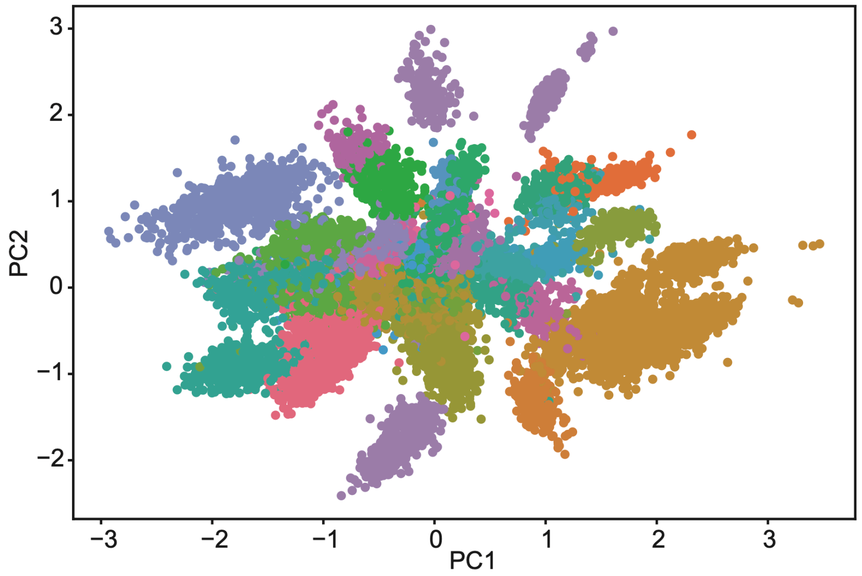
Does the model learn?

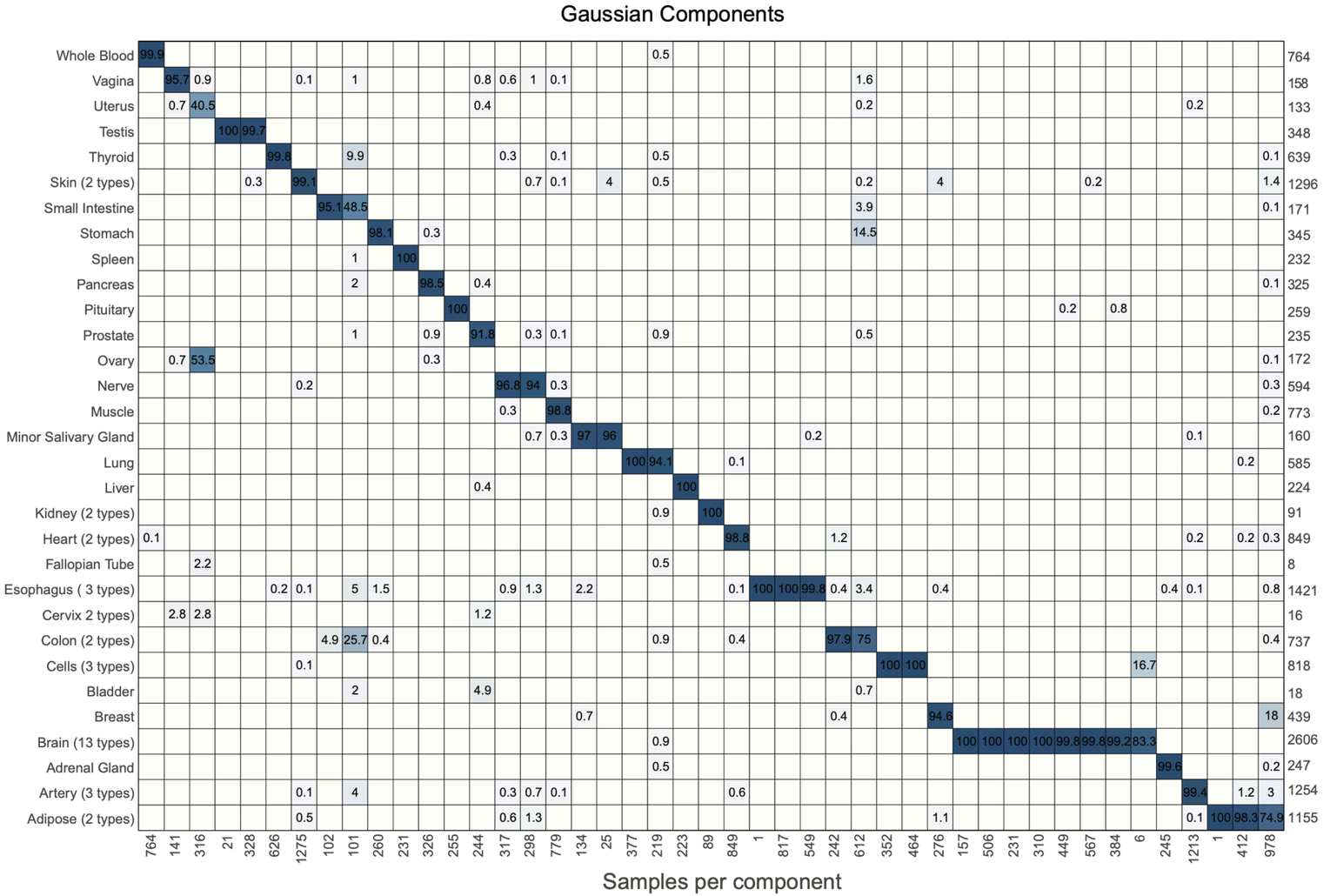




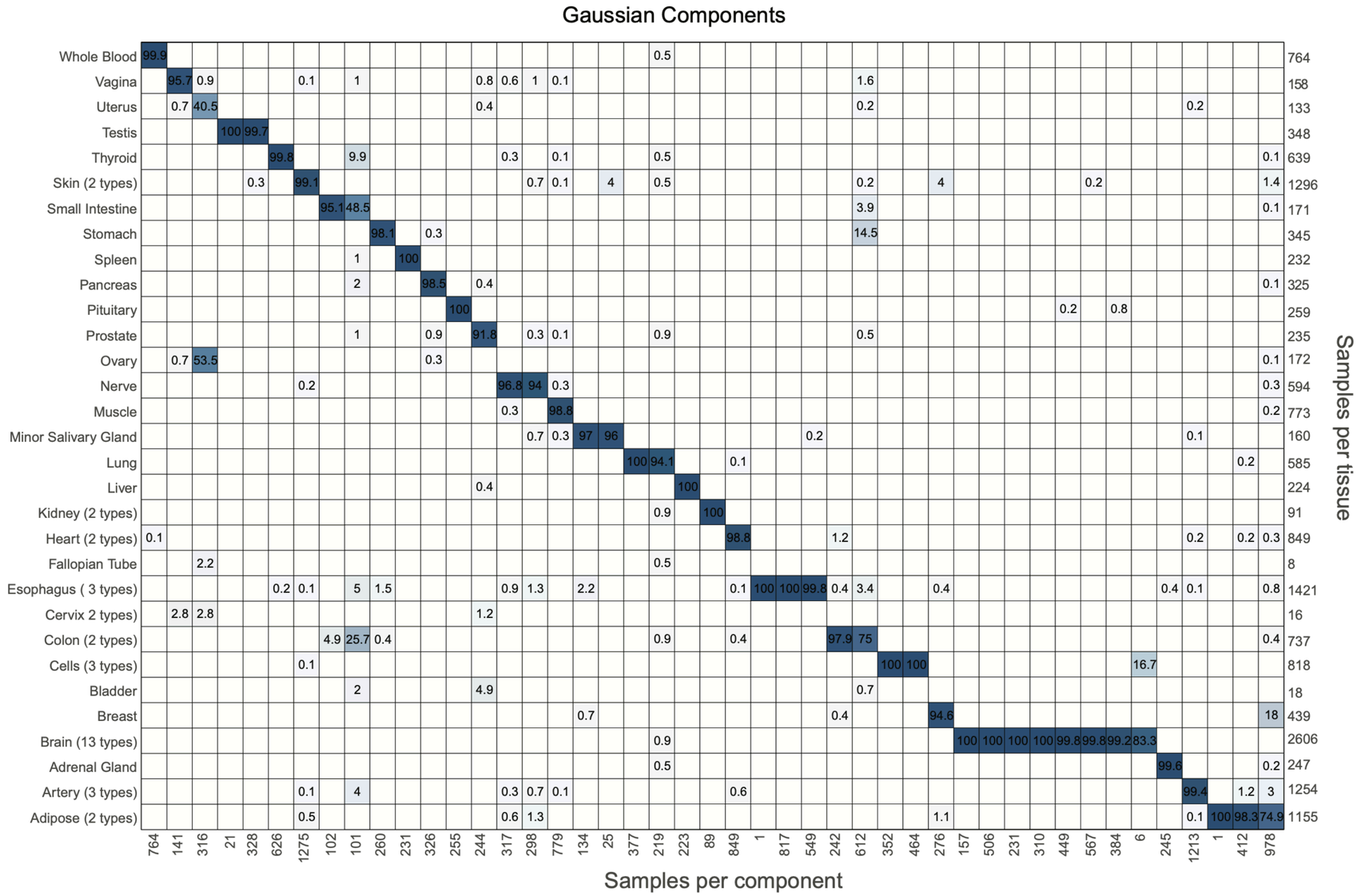
Does the model learn?







Does the model learn?

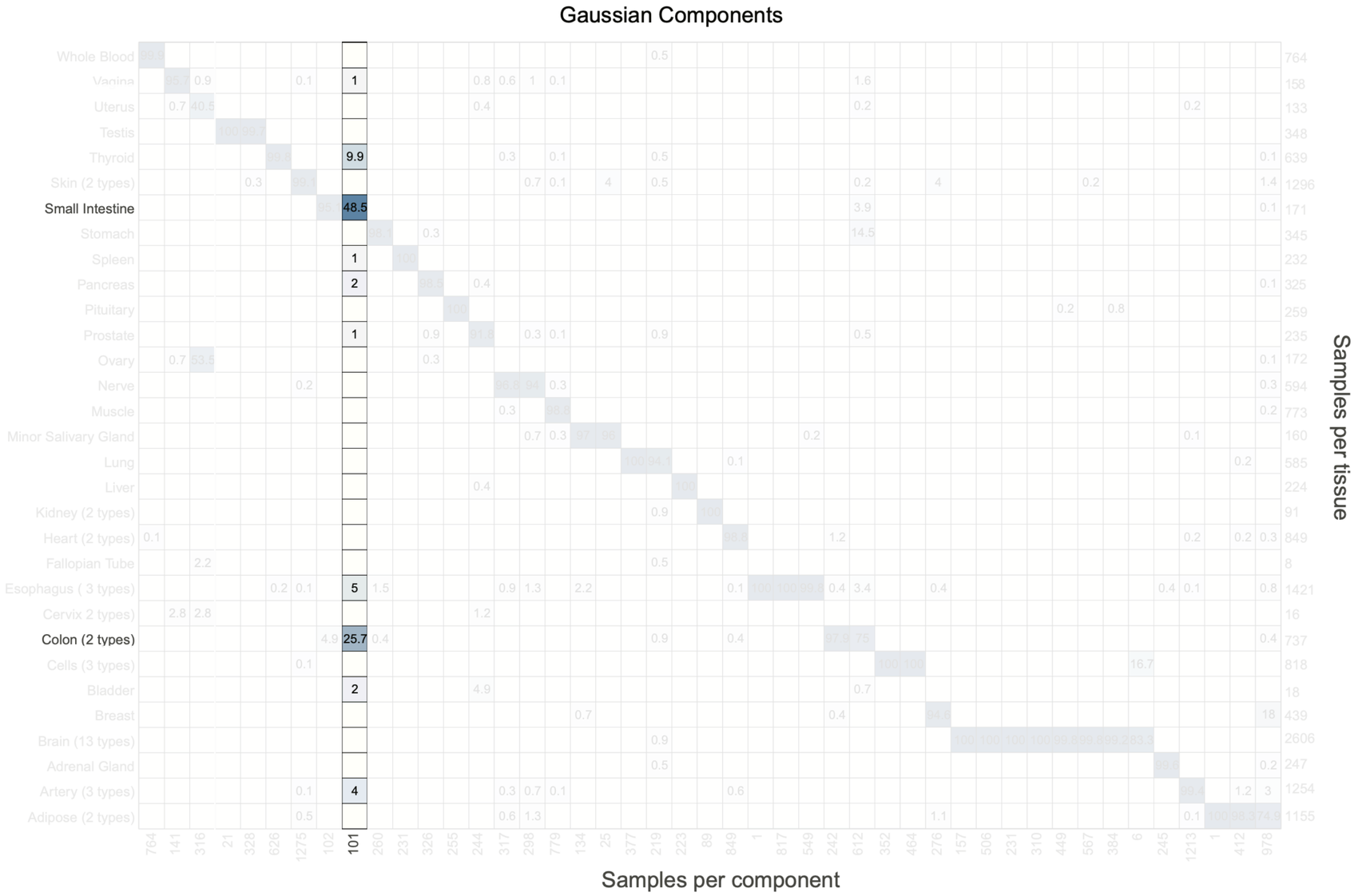


- Single tissue dominates most of components

Does the model learn?

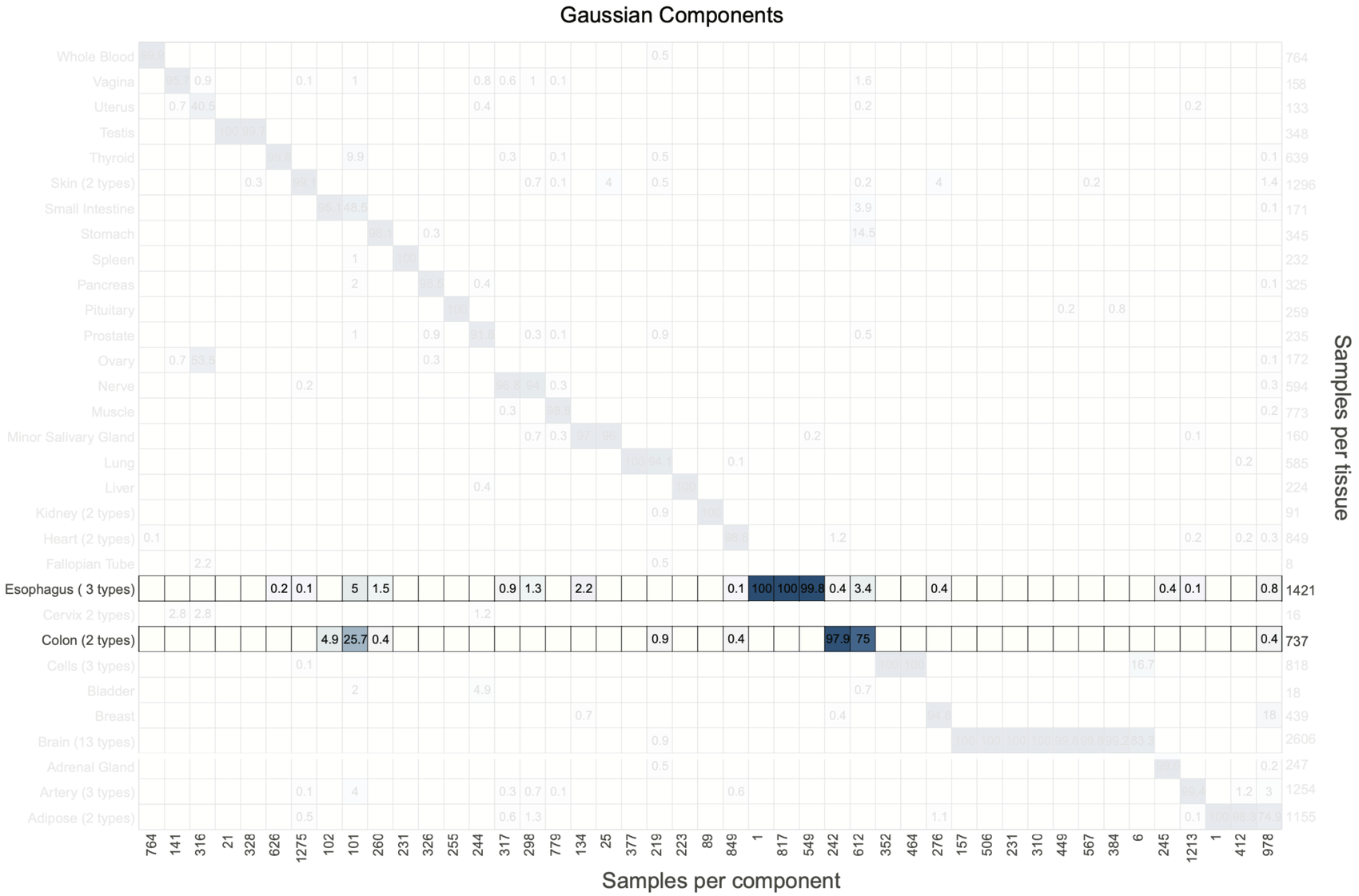
- Single tissue dominates most of components

Does the model learn?



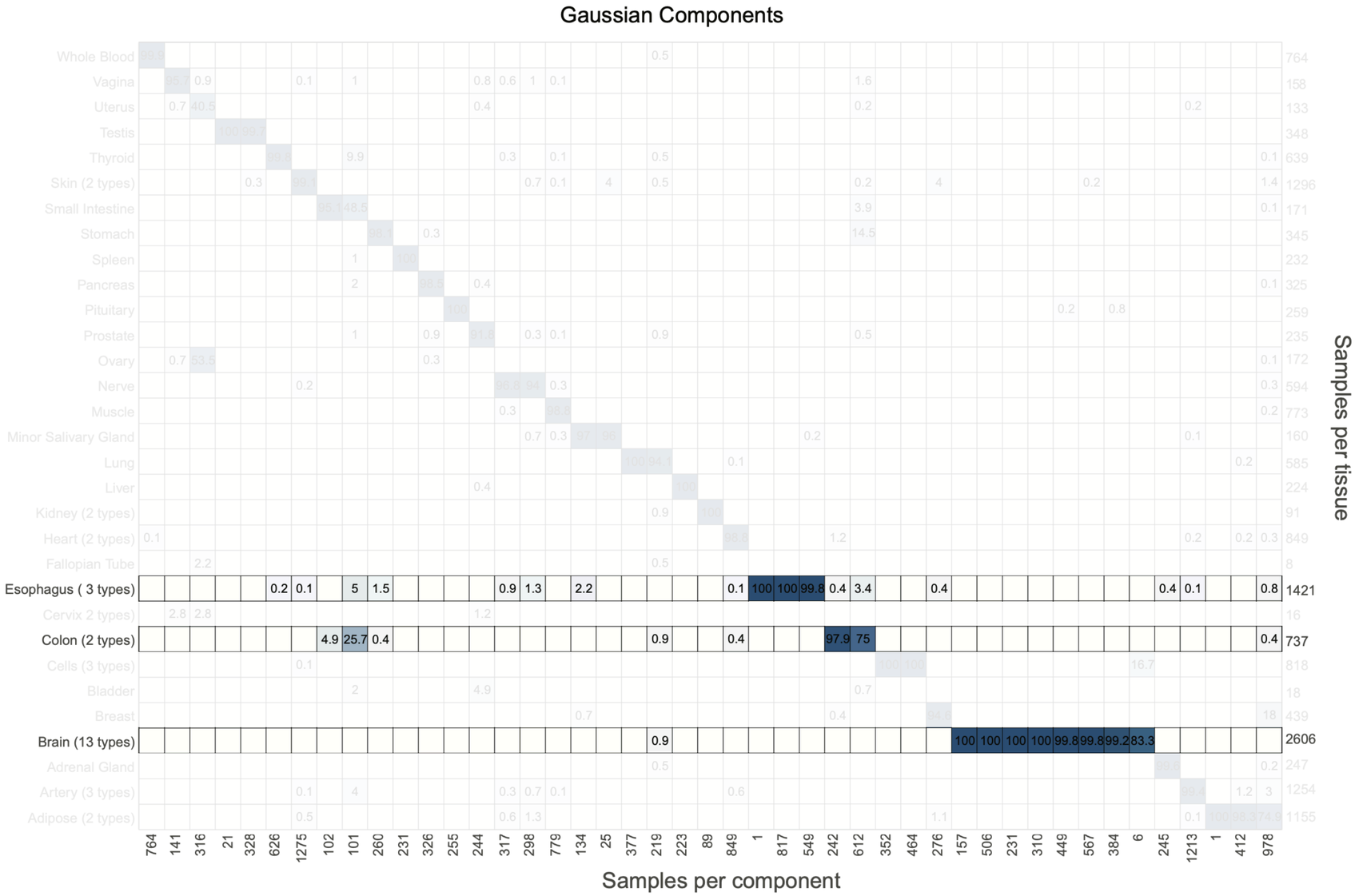
- Single tissue dominates most of components

Does the model learn?



- Single tissue dominates most of components
- Tissues are divided according to biological functions

Does the model learn?



- Single tissue dominates most of components
- Tissues are divided according to biological functions

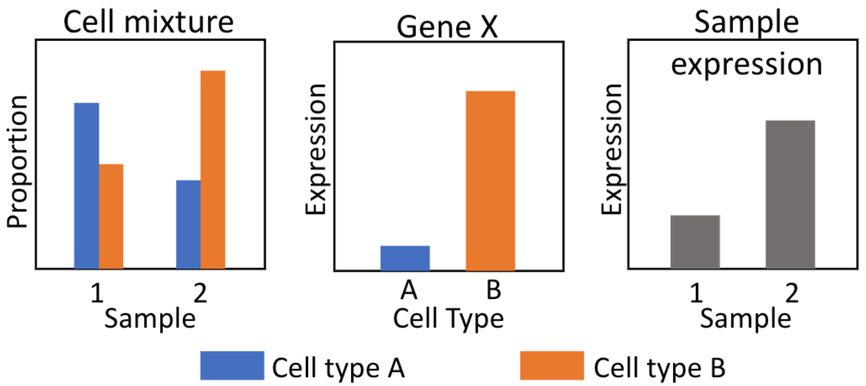
this is very nice but what can you do with it?

What can we use the model for?

- Project new samples into the model:
  - Create *in silico samples* **Often we have no good controls!**

What can we use the model for?

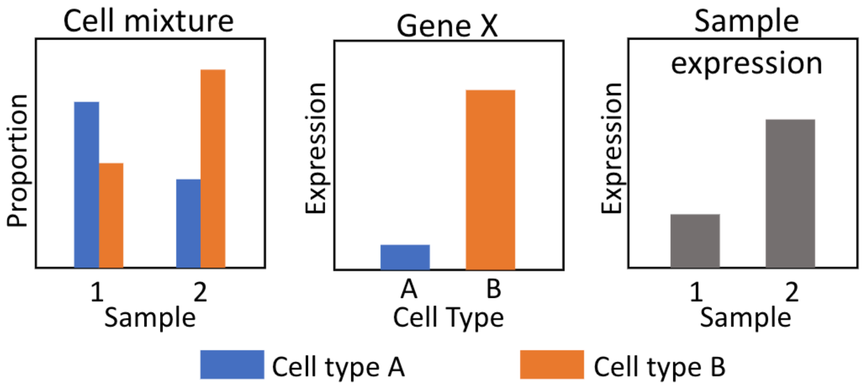
- Project new samples into the model:
  - Create *in silico samples* **Often we have no good controls!**
    - In the clinic:
      - Samples from other individuals are genetically different
      - In cancer, adjacent normal samples may resemble disease
    - Even if you have a good control
      - Likely to have cell-to-cell differences



      - This leads to biased differential expression estimates

What can we use the model for?

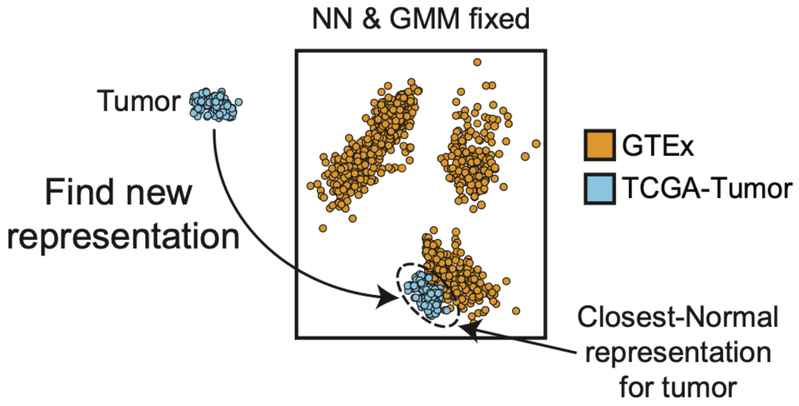
- Project new samples into the model:
  - Create *in silico samples* **Often we have no good controls!**
    - In the clinic:
      - Samples from other individuals are genetically different
      - In cancer, adjacent normal samples may resemble disease
    - Even if you have a good control
      - Likely to have cell-to-cell differences



      - This leads to biased differential expression estimates
  - Find comparison tissues and tissues of origin for cancer samples

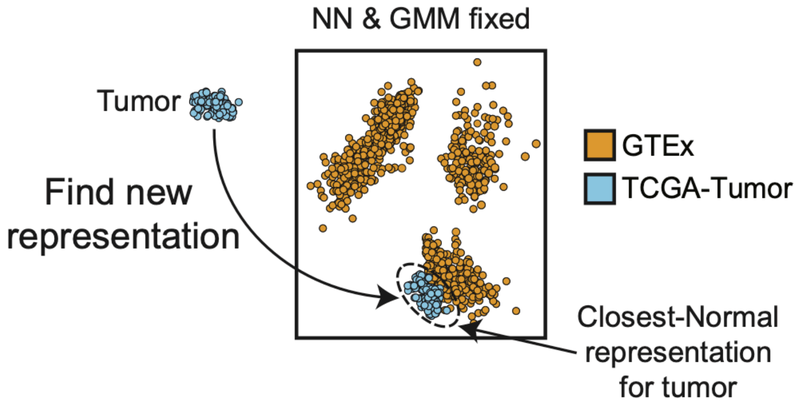
Finding the ideal comparison sample

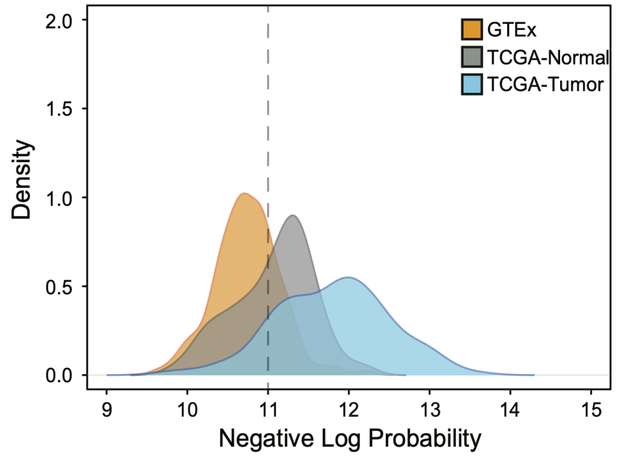
What is a good comparison sample?



- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

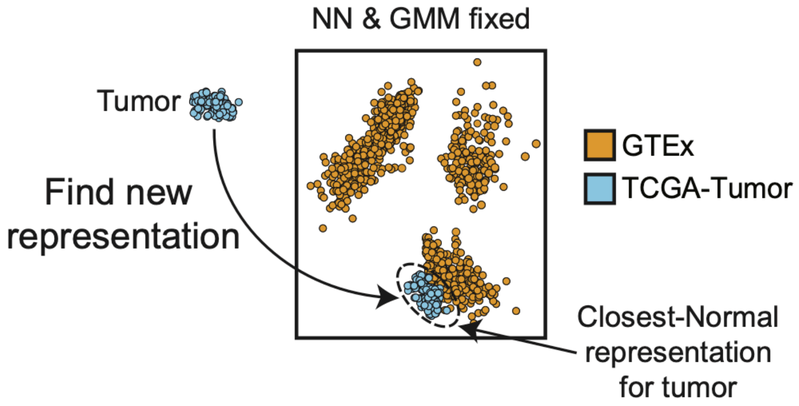
What is a good comparison sample?

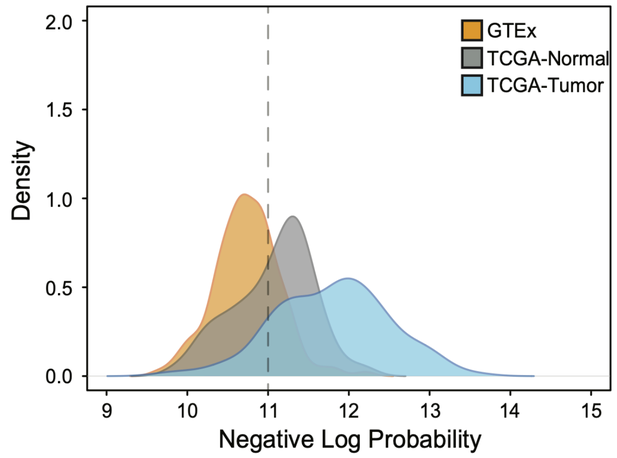




- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

What is a good comparison sample?

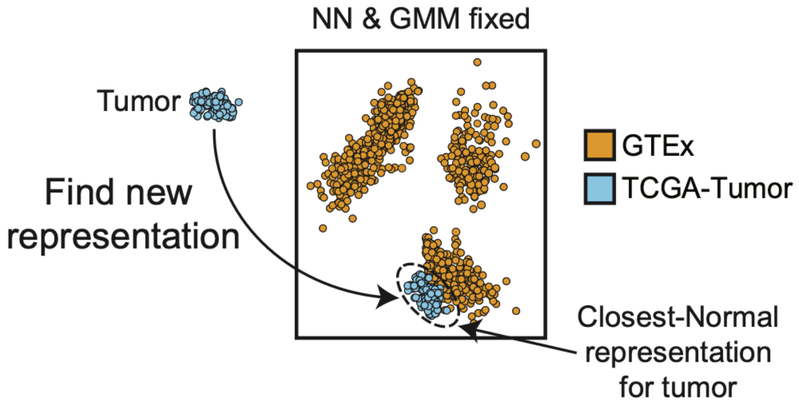




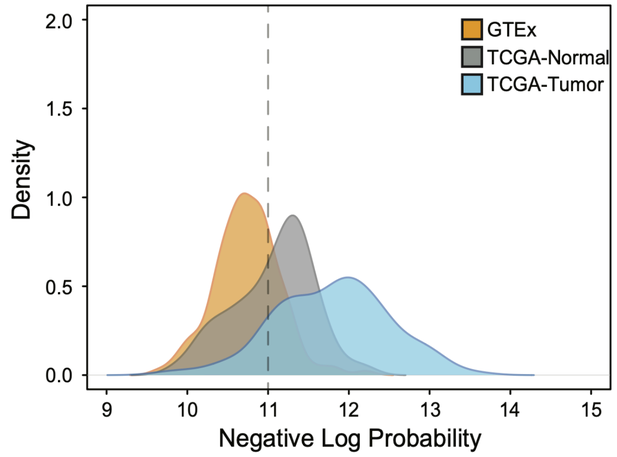
- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

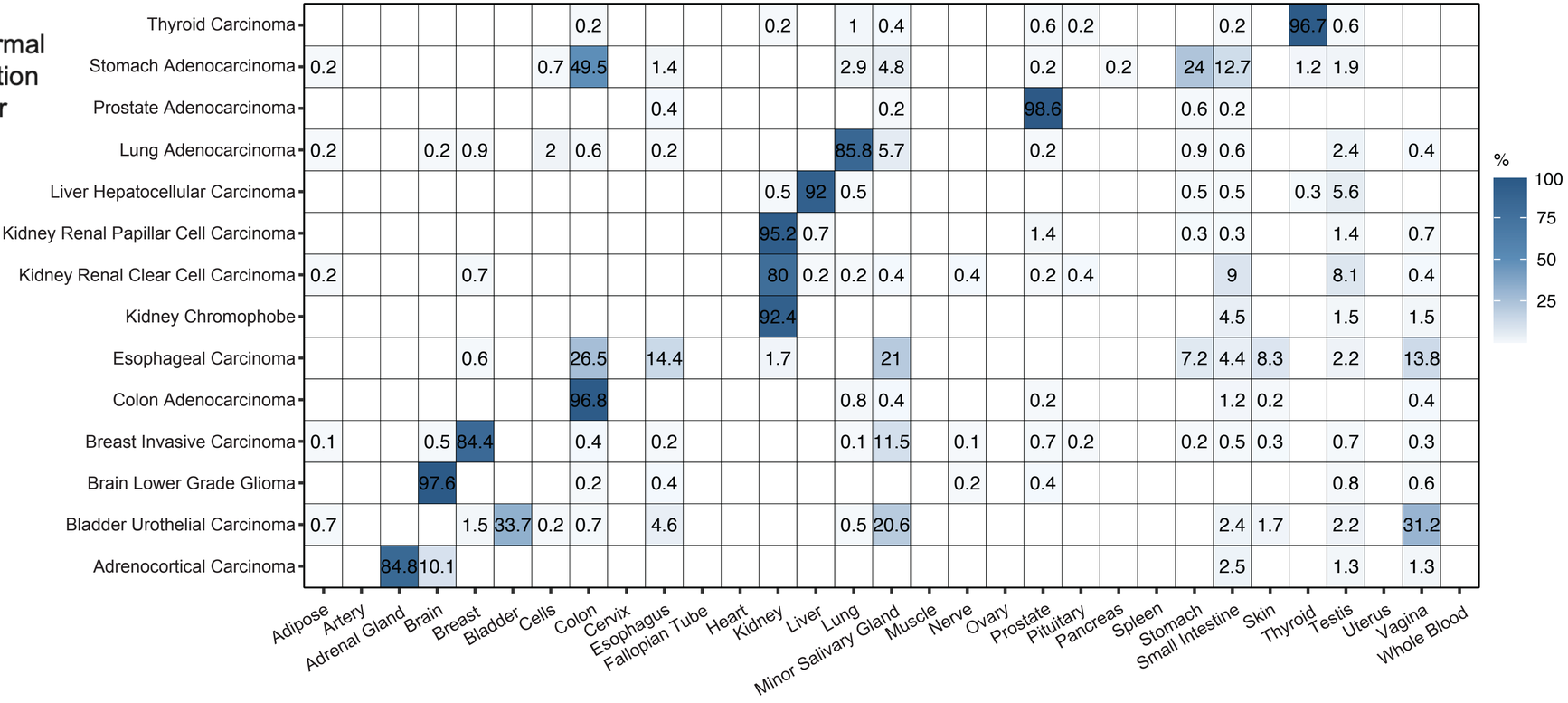
Thyroid Carcinoma
Stomach Adenocarcinoma
Prostate Adenocarcinoma
Lung Adenocarcinoma
Liver Hepatocellular Carcinoma
Kidney Renal Papillar Cell Carcinoma
Kidney Renal Clear Cell Carcinoma
Kidney Chromophobe
Esophageal Carcinoma
Colon Adenocarcinoma
Breast Invasive Carcinoma
Brain Lower Grade Glioma
Bladder Urothelial Carcinoma
Adrenocortical Carcinoma

What is a good comparison sample?

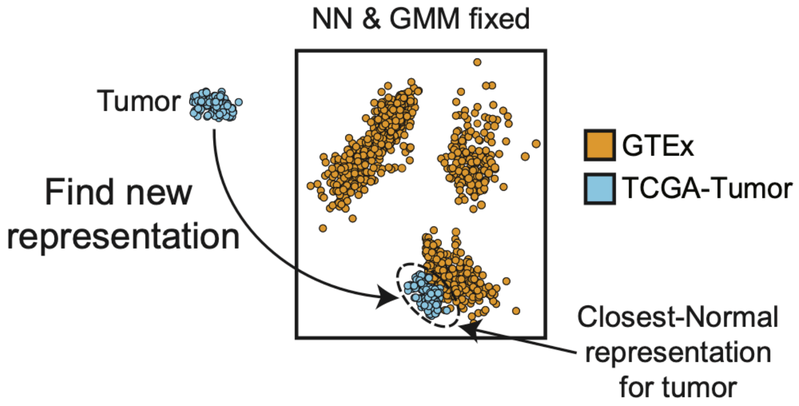


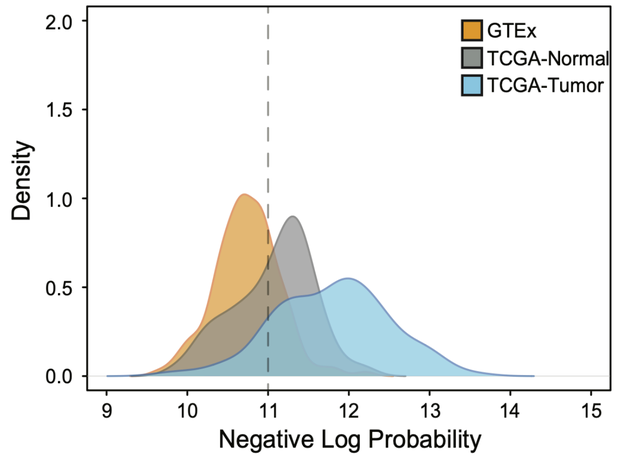
- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?



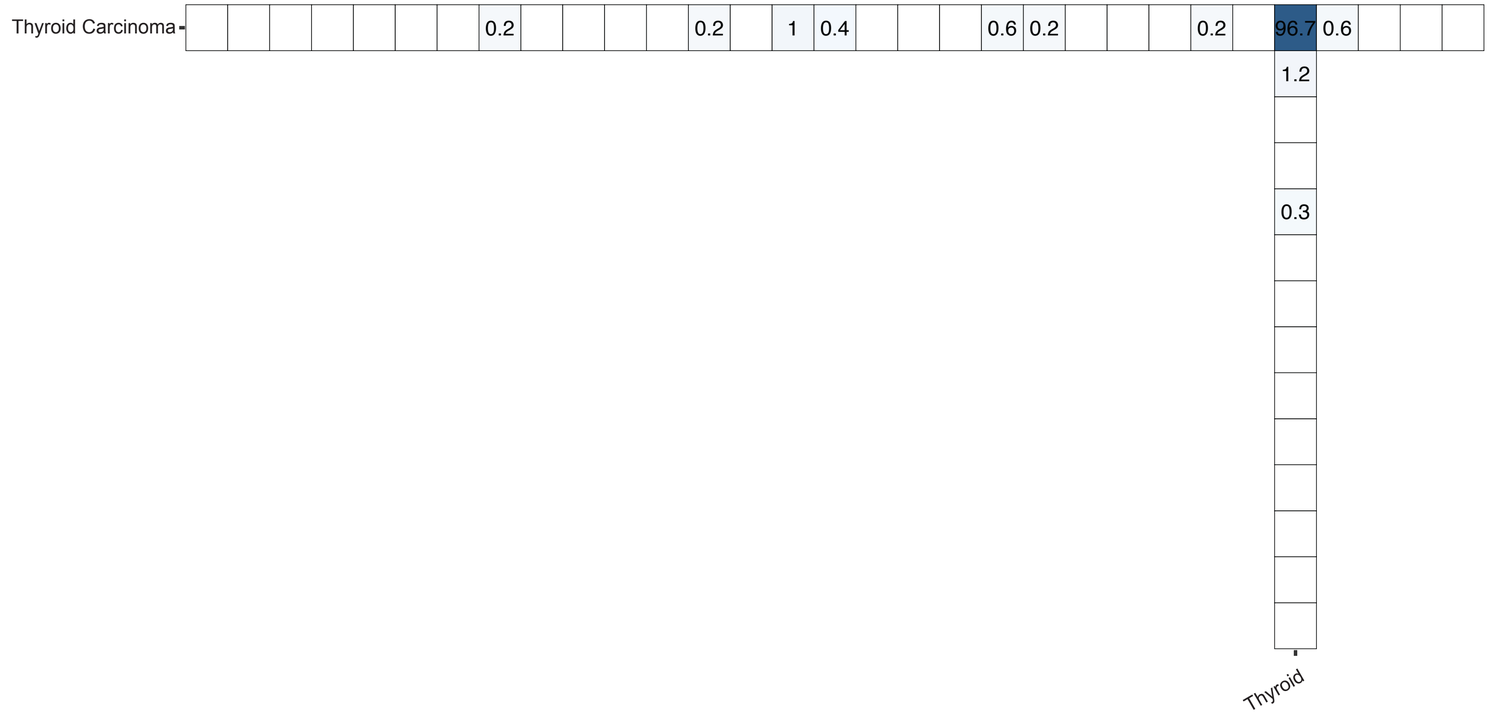


What is a good comparison sample?

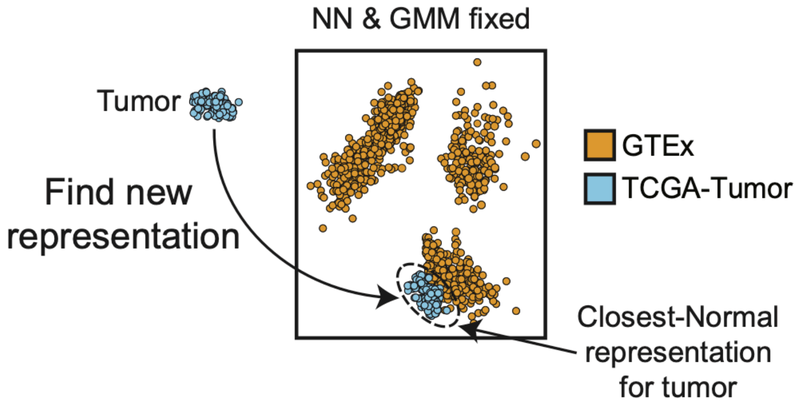


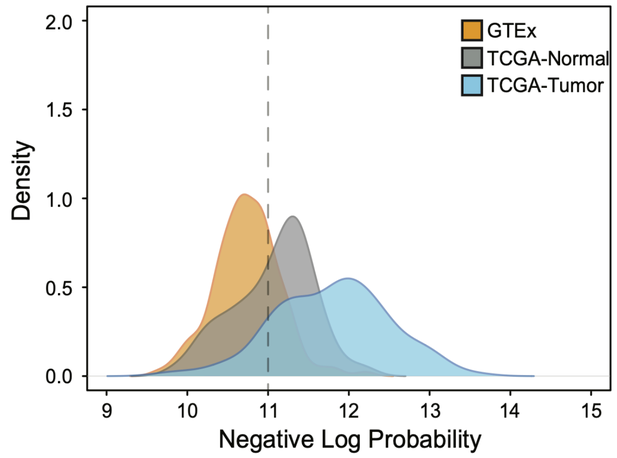


- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

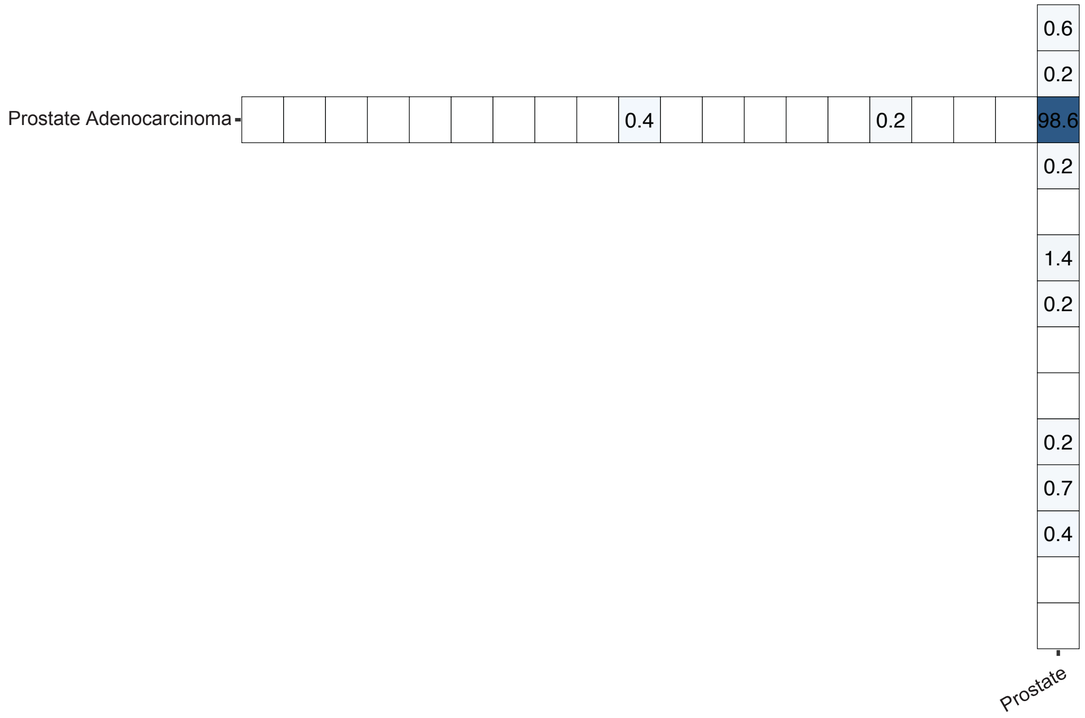


What is a good comparison sample?

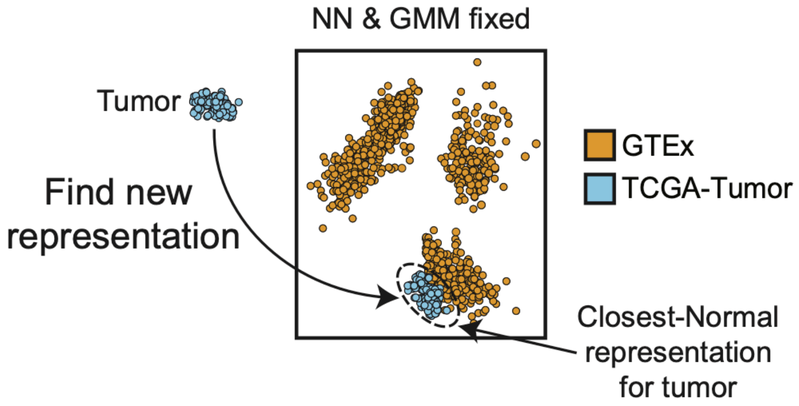


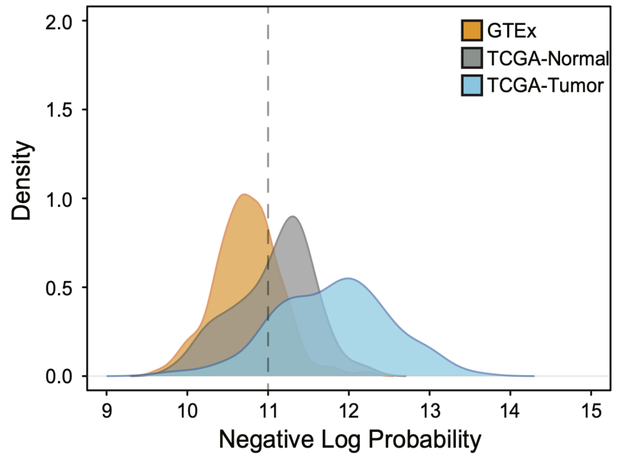


- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

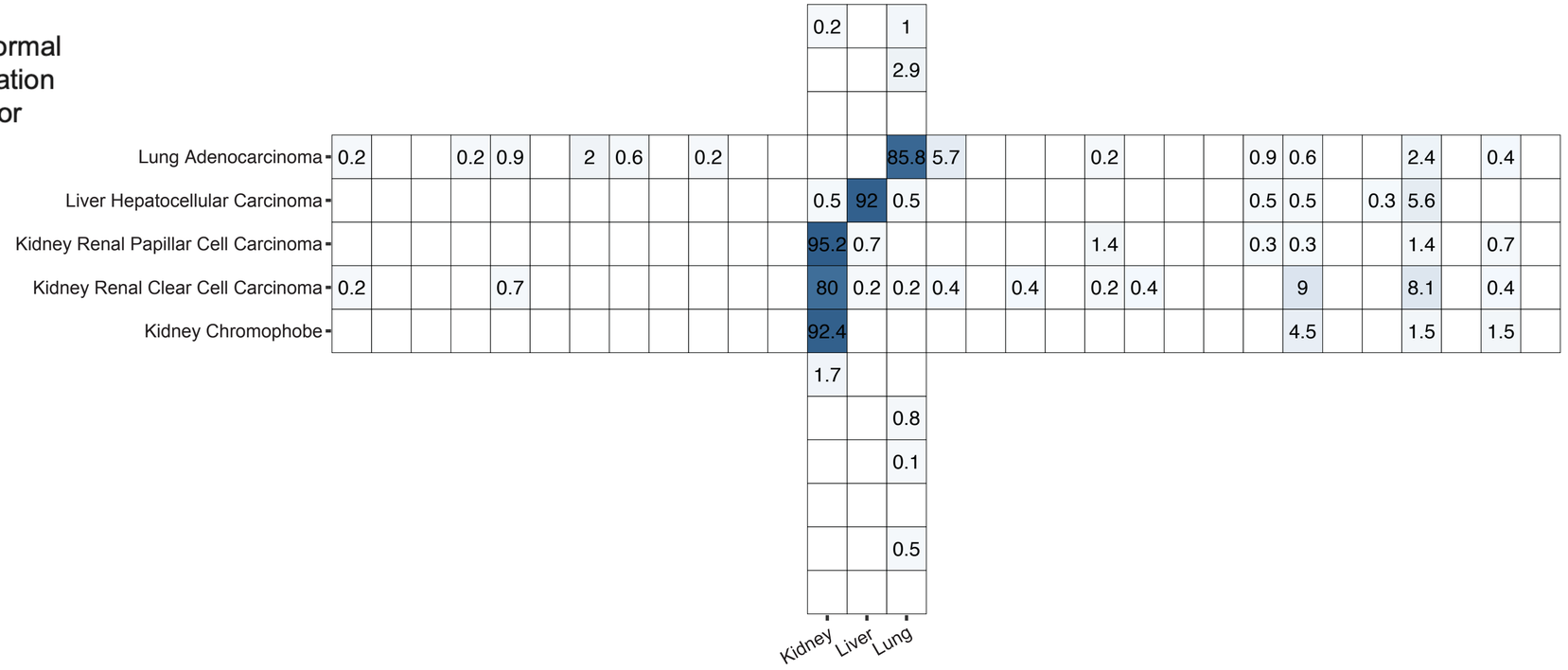


What is a good comparison sample?

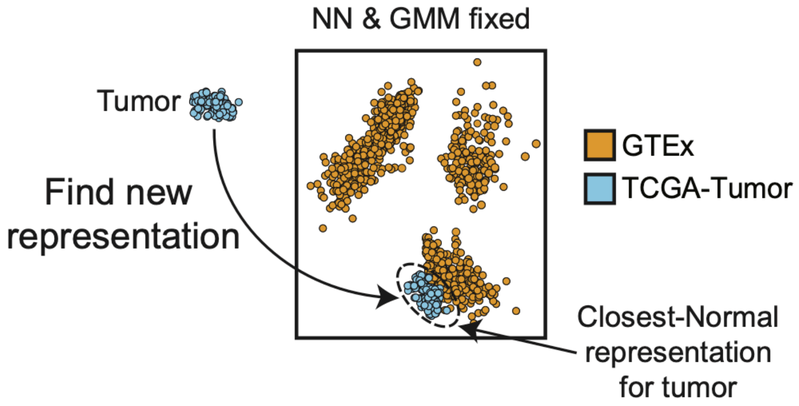




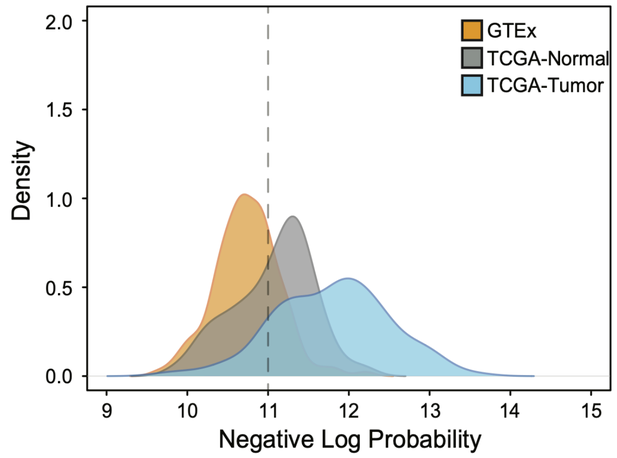
- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?

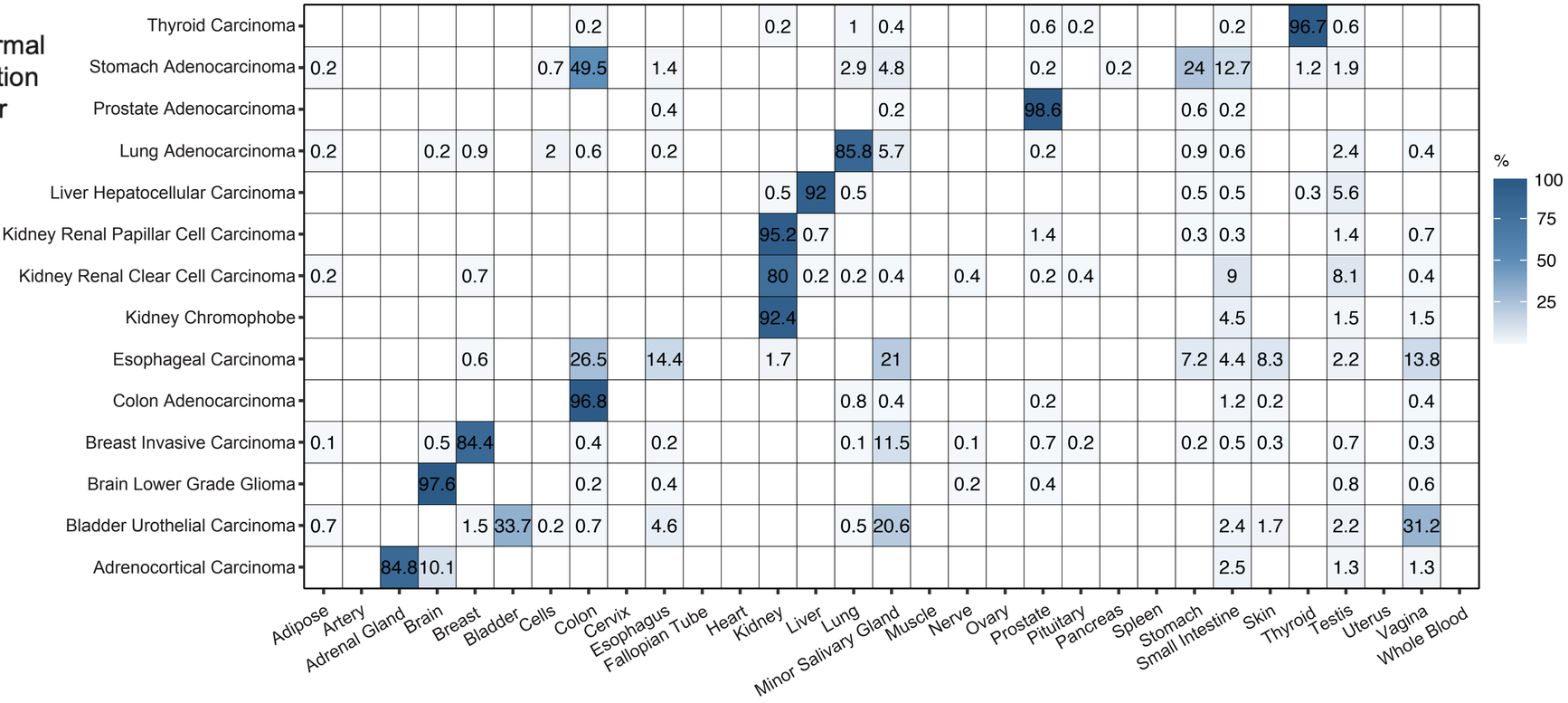


What is a good comparison sample?



- Can the model select the ideal comparison sample?
  - Find closest-normal representations for cancer genome atlas samples
- Can the model detect the tissue of origin?



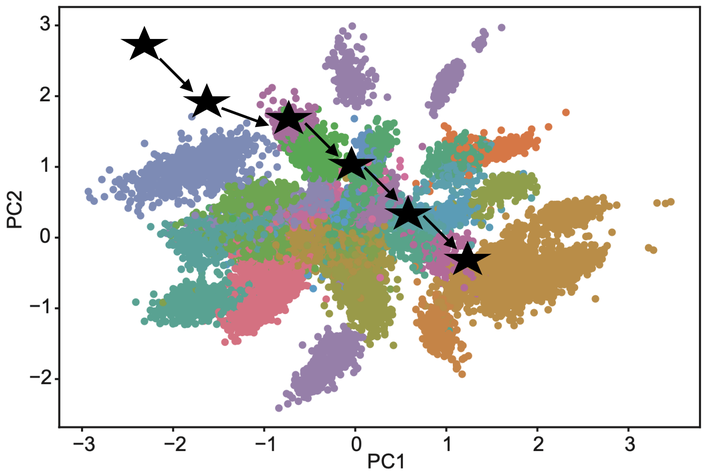


The model finds adequate *in silico* comparison samples

A new way of finding differentially expressed genes?

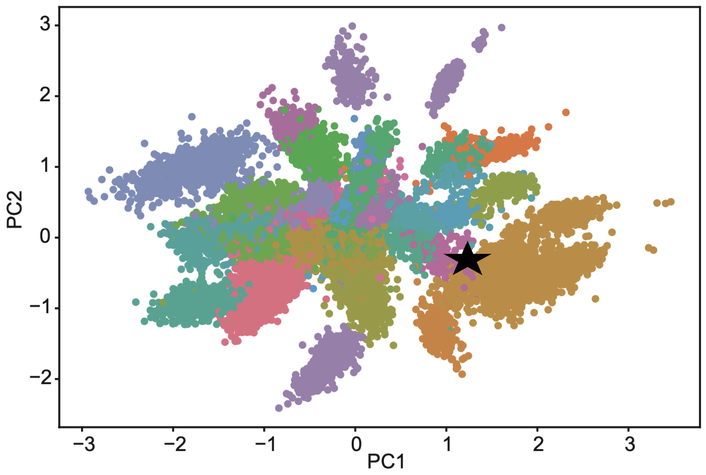
- Standard set-up: *N* cases vs *N* controls
  - Samples and cases are hard to find.
  - Analysis suffers from low sample sizes
  - Selection of good controls is critical
- We propose: Deep generative model vs single patient
  - Train a generative model on large-scale bulk RNA-seq
  - *Generate* the control you need for you disease sample
  - Allows for personalised analysis in an *N-of-one* manner

A new way of finding differentially expressed genes?



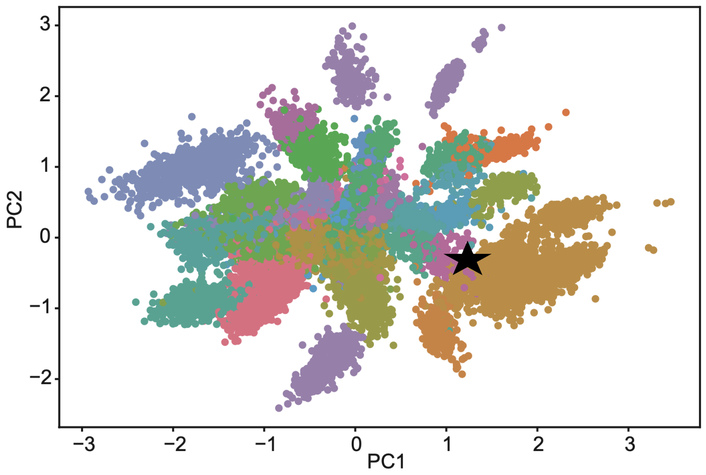
★ Most likely representation

A new way of finding differentially expressed genes?

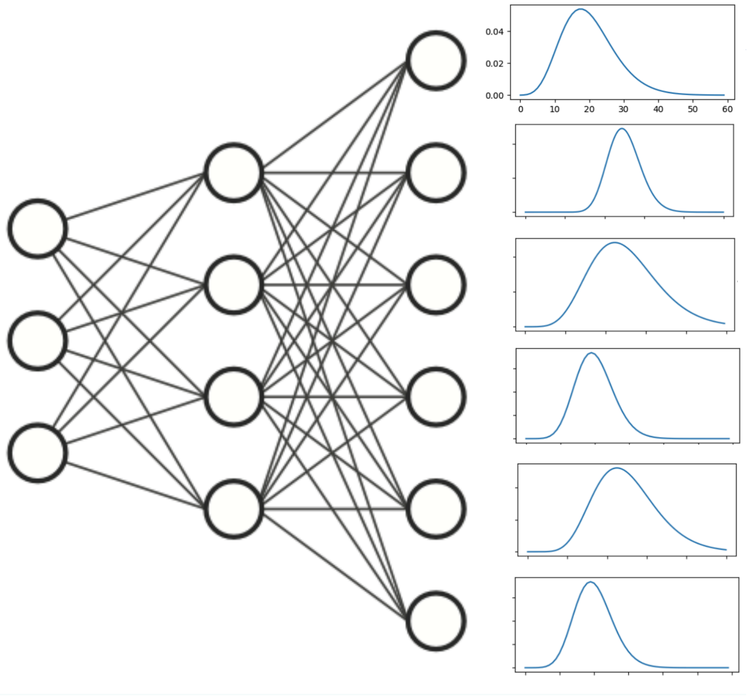


★ Most likely representation

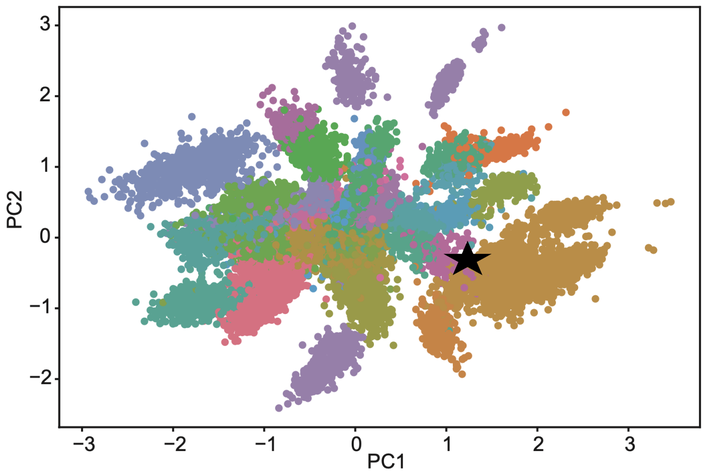
A new way of finding differentially expressed genes?



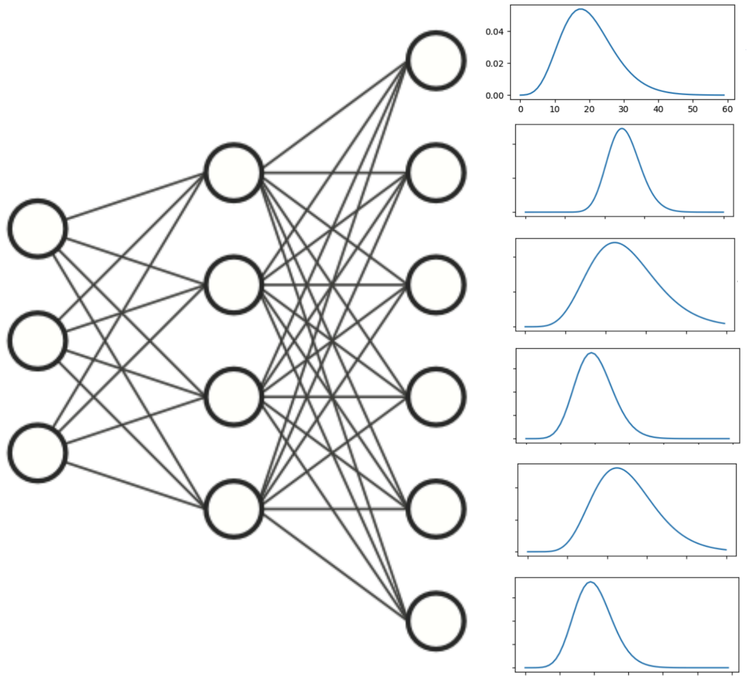
★ Most likely representation



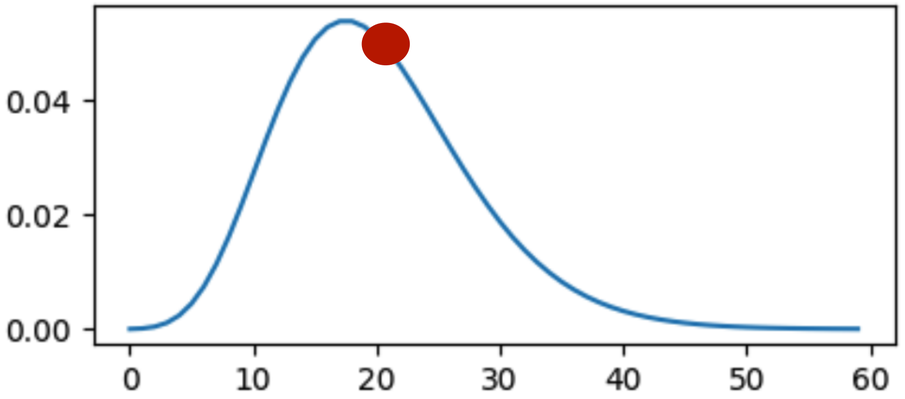
A new way of finding differentially expressed genes?



★ Most likely representation

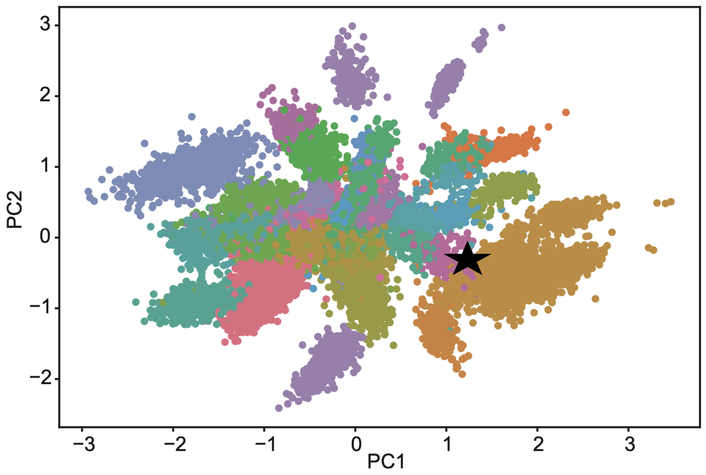


Gene of interest

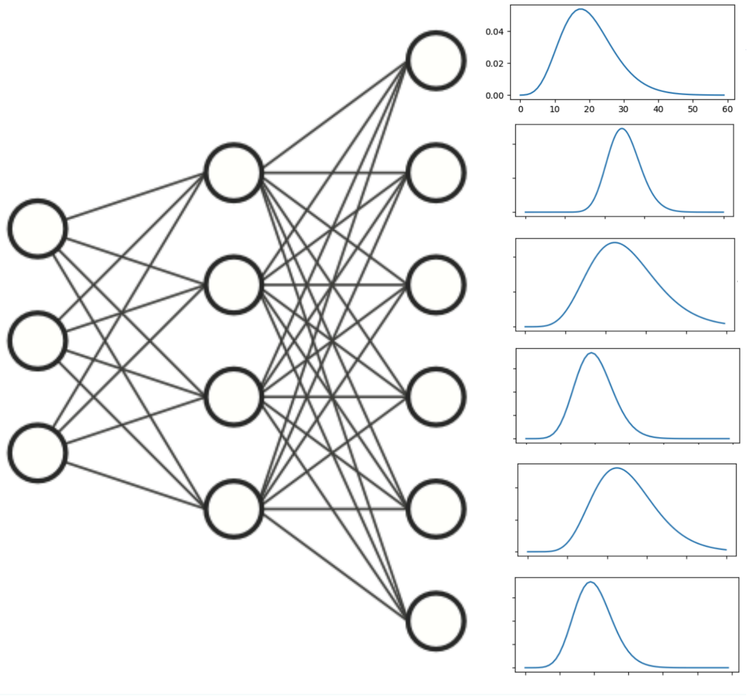


Observed count

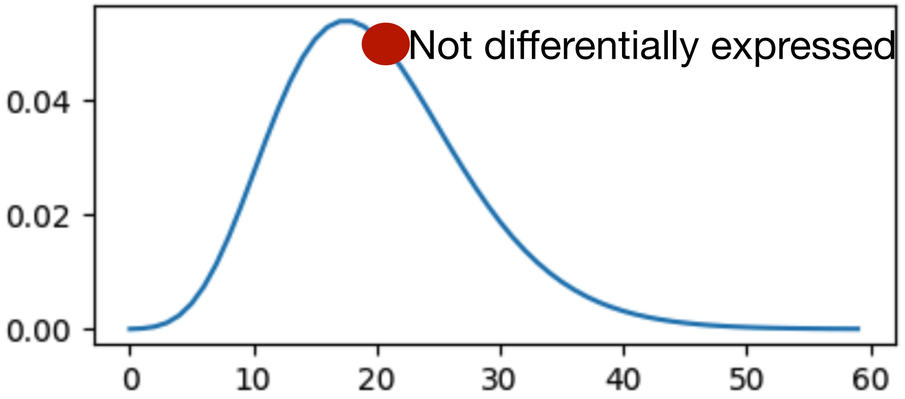
A new way of finding differentially expressed genes?



★ Most likely representation

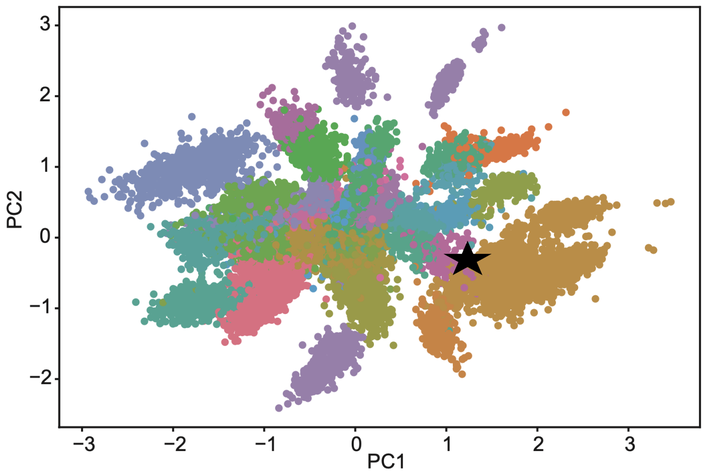


Gene of interest

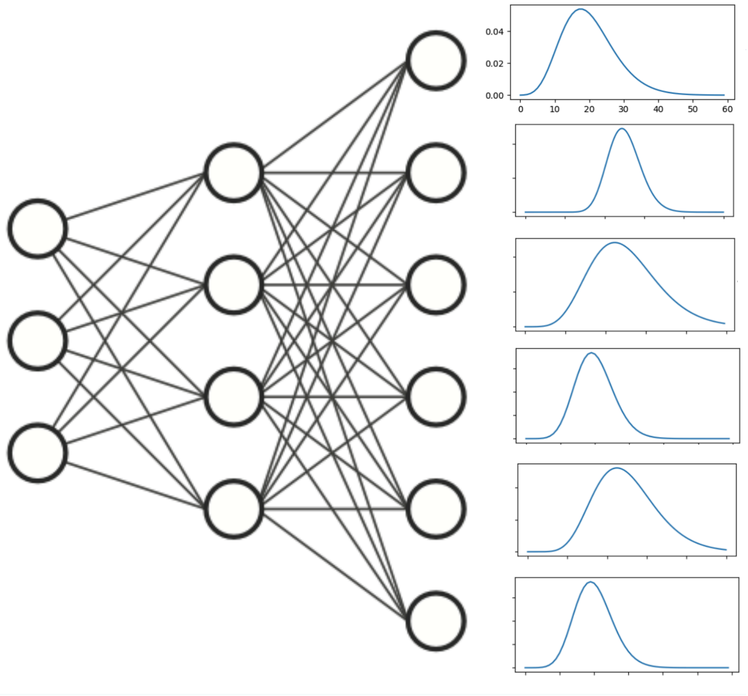


Observed count

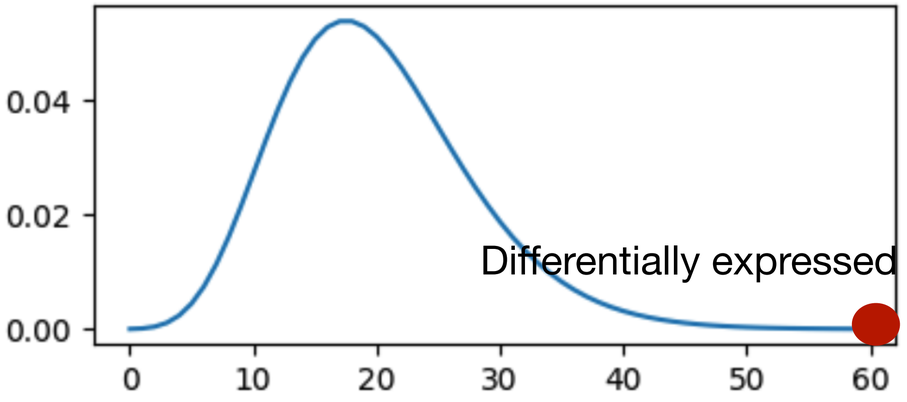
A new way of finding differentially expressed genes?



★ Most likely representation



Gene of interest



Observed count

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

Methods

- DEseq2
- DGD

Data

Healthy vs healthy:
GTEx breast

Healthy vs cancer:
GTEx breast
TCGA-Tumor

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

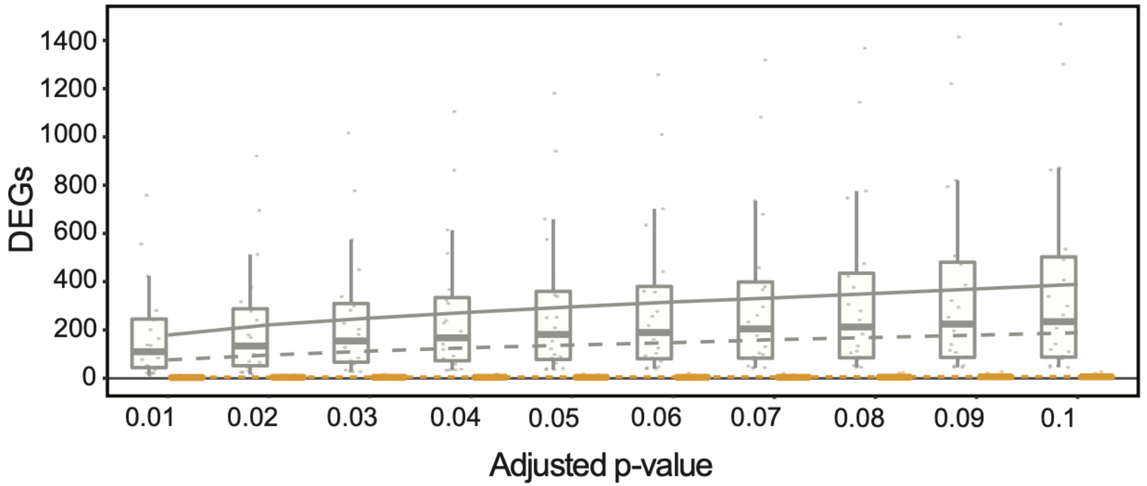
Methods

DEseq2
DGD

Data

Healthy vs healthy:
GTEx breast

Healthy vs cancer:
GTEx breast
TCGA-Tumor



Healthy vs healthy:
Proxy to specificity

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

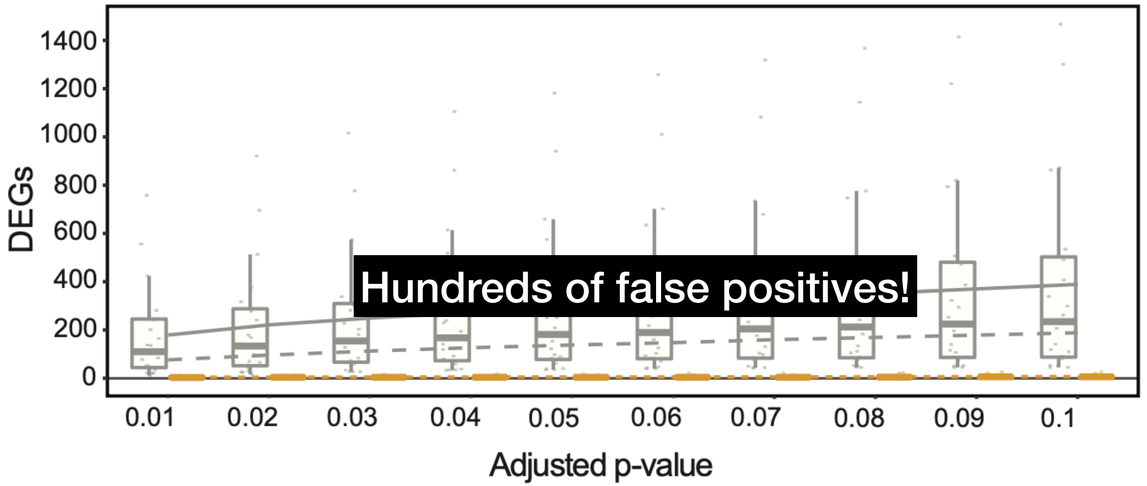
Methods

DEseq2
DGD

Data

Healthy vs healthy:
GTEx breast

Healthy vs cancer:
GTEx breast
TCGA-Tumor



Healthy vs healthy:
Proxy to specificity

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

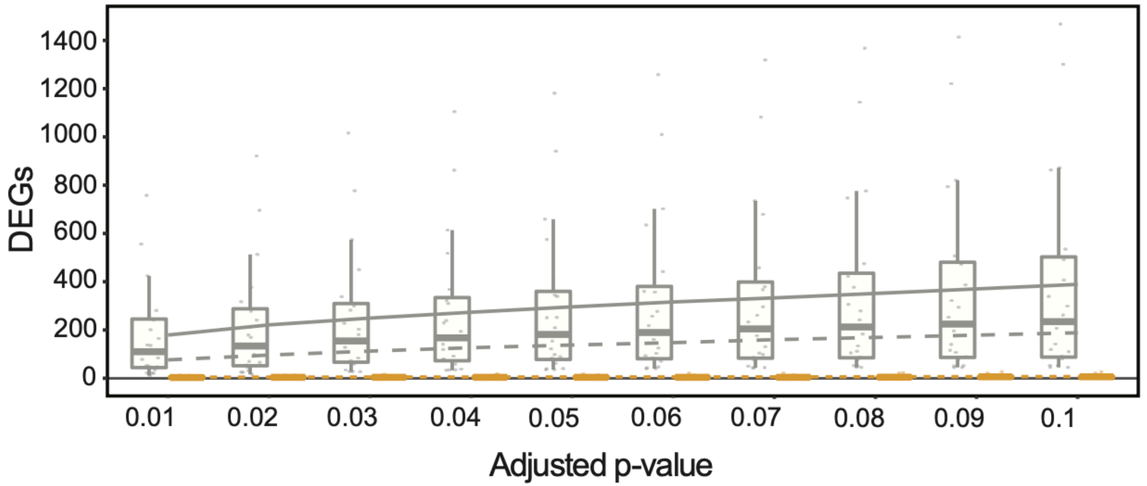
Methods

DEseq2
DGD

Data

Healthy vs healthy:
GTEx breast

Healthy vs cancer:
GTEx breast
TCGA-Tumor



Healthy vs healthy:
Proxy to specificity

Healthy vs cancer on marker genes
Proxy to sensitivity

Marker gene sets
DriverDB
PAM50

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

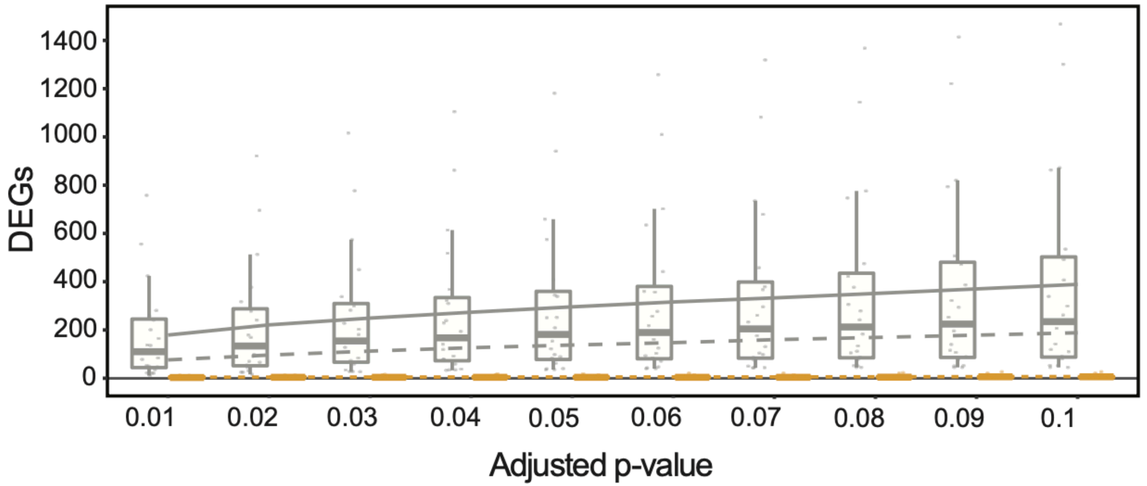
Methods

DEseq2
DGD

Data

Healthy vs healthy:
GTEx breast

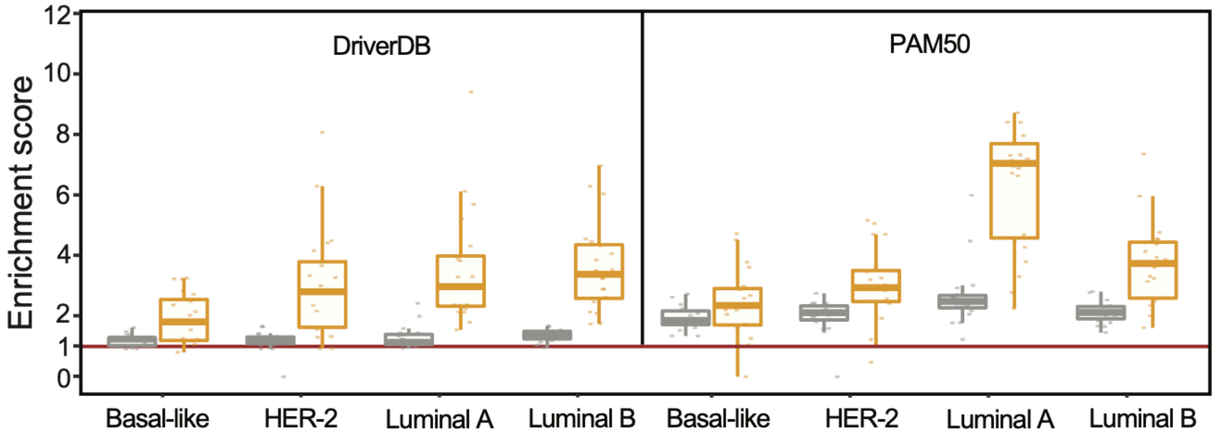
Healthy vs cancer:
GTEx breast
TCGA-Tumor



Healthy vs healthy:
Proxy to specificity

Healthy vs cancer on marker genes
Proxy to sensitivity

Marker gene sets
DriverDB
PAM50



Score: $\frac{\text{Fraction of marker significant genes}}{\text{Fraction of all significant genes}}$

Can we revisit gene expression?

Experiment:
1-Select a **single** cancer sample
2-Compare to 5 controls or DGD
3-Repeat 30 times

Calling differential expression:
Negative binomial test between closest-normal sample and tumor sample

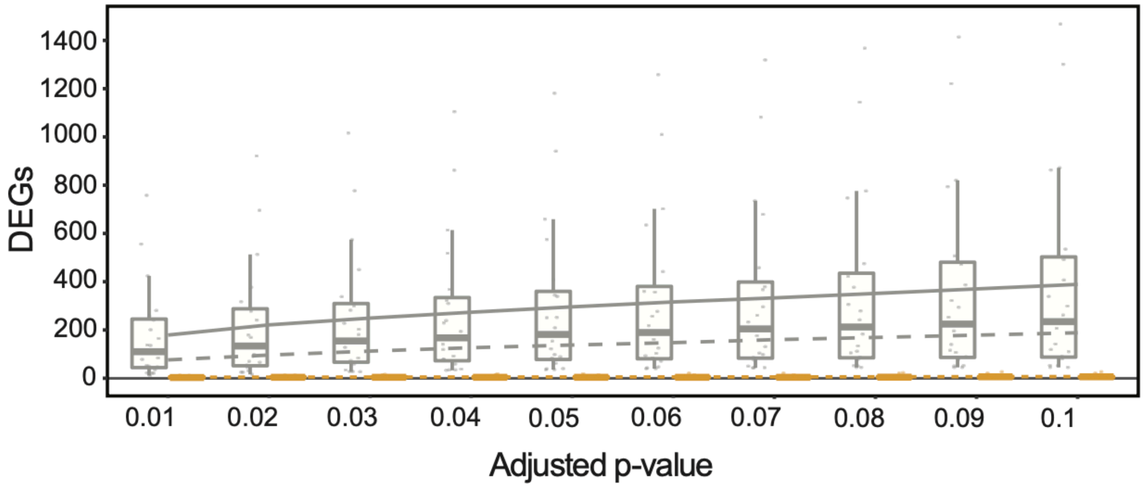
Methods

DEseq2
DGD

Data

Healthy vs healthy:
GTEx breast

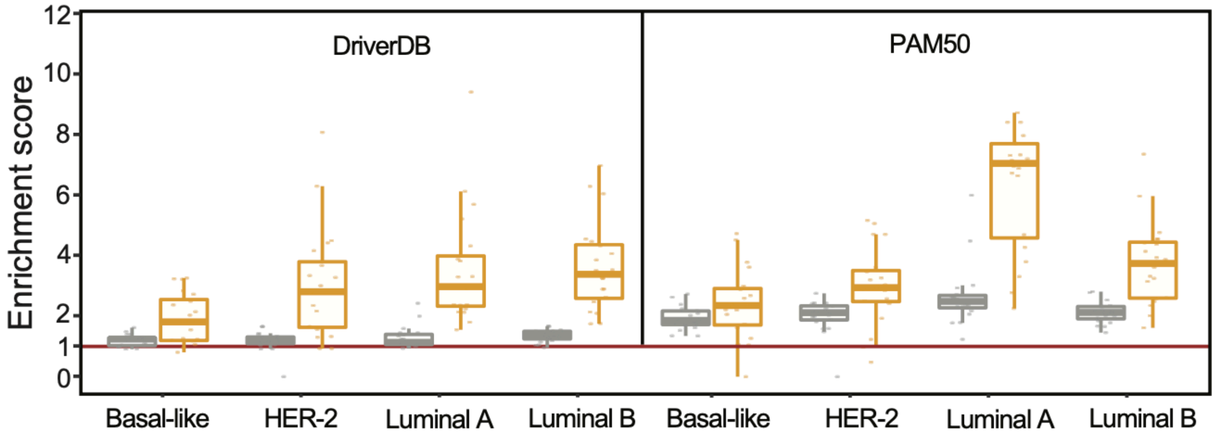
Healthy vs cancer:
GTEx breast
TCGA-Tumor



Healthy vs healthy:
Proxy to specificity

Healthy vs cancer on marker genes
Proxy to sensitivity

Marker gene sets
DriverDB
PAM50



Score: $\frac{\text{Fraction of marker significant genes}}{\text{Fraction of all significant genes}}$

DGD has a higher sensitivity, while keeping false positives low

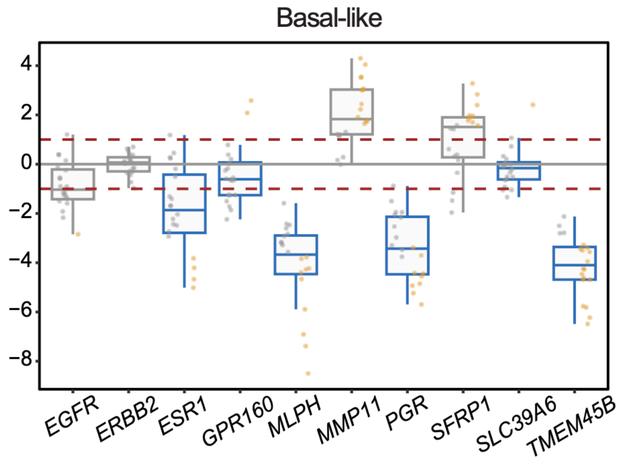
Zooming in breast cancer

Can we find the well known breast cancer genes?

Do we correctly guess expression trends? (Up/down)

P-value < 0.05
P-value > 0.05

Upregulated
Normal expression
Downregulated



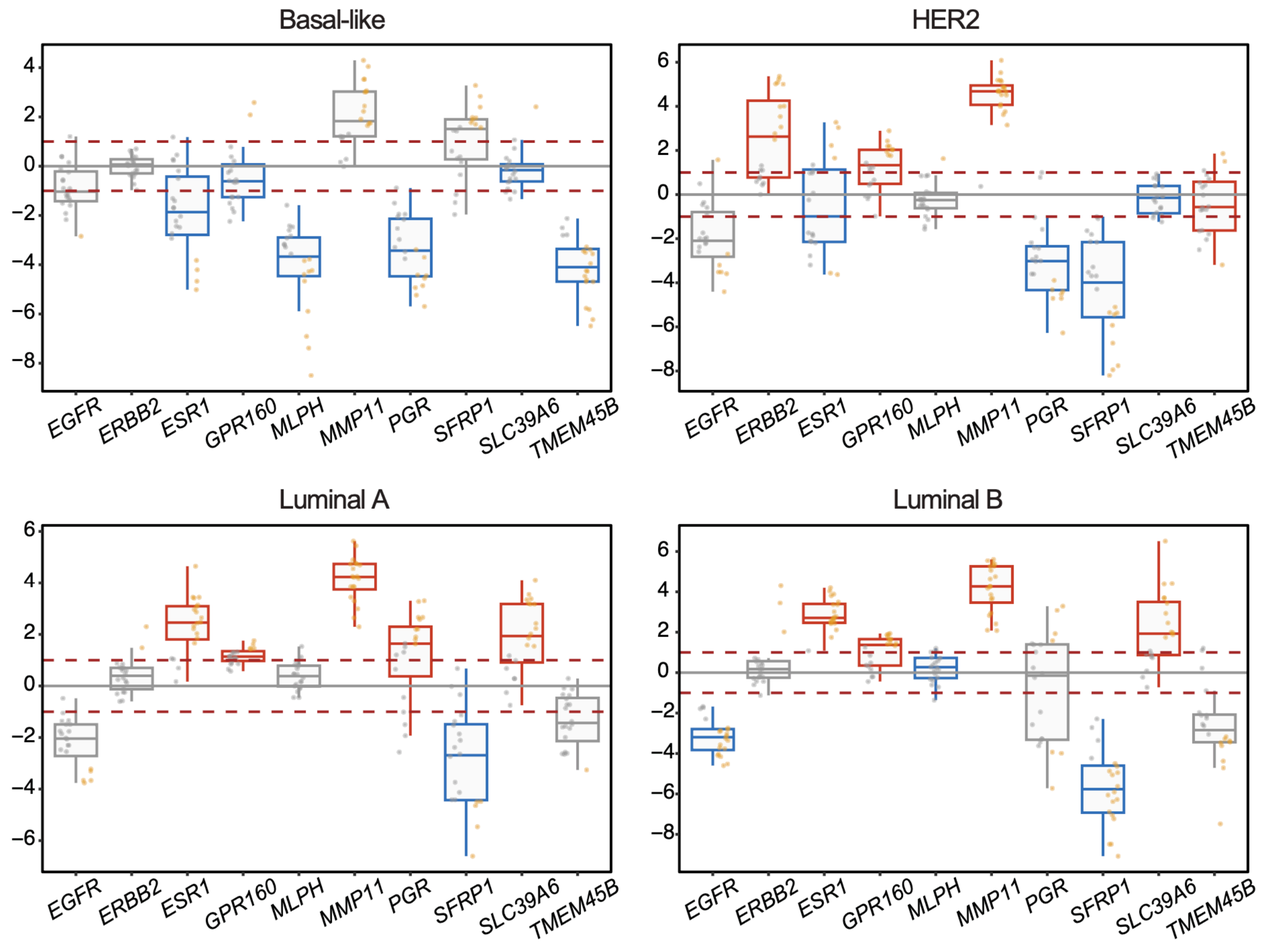
Zooming in breast cancer

Can we find the well known breast cancer genes?

Do we correctly guess expression trends? (Up/down)

P-value < 0.05
P-value > 0.05

Upregulated
Normal expression
Downregulated



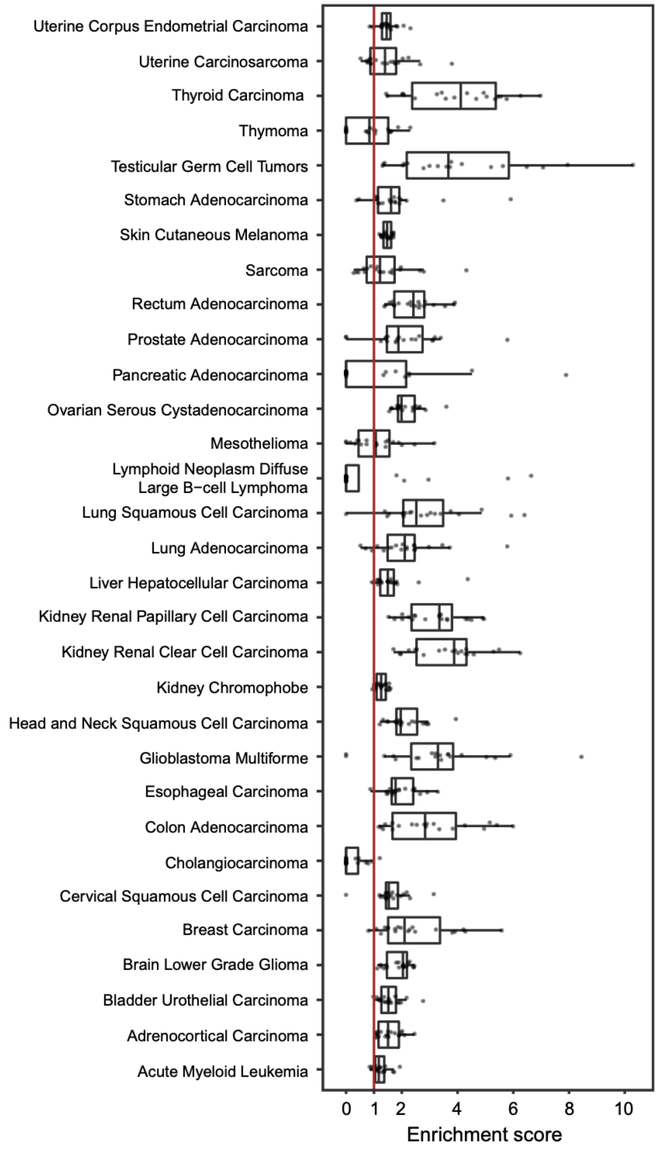
DGD finds the main patterns of gene expression across the subtypes

Do we find enrichment across cancers?

- Good performance in breast cancer
  - Better sensitivity than DEseq2
  - Dramatically low false positives
  - Very few called genes, facilitating interpretation
- How is the performance in other cancers?
  - Extend our analysis across the cancer genome atlas

Do we find enrichment across cancers?

- Good performance in breast cancer
  - Better sensitivity than DEseq2
  - Dramatically low false positives
  - Very few called genes, facilitating interpretation
- How is the performance in other cancers?
  - Extend our analysis across the cancer genome atlas



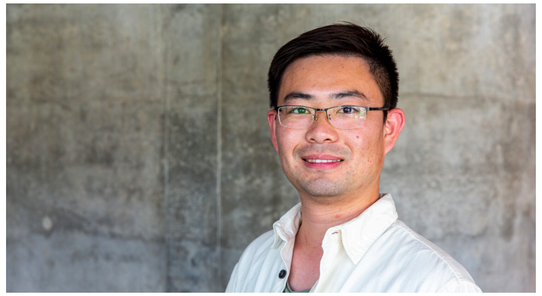
Summary and conclusions

- DGD beats the standards.
  - Note: with only 2 fully-connected layers
- Keeps the gene “haystack” small
- Stop using controls!
  - Are healthy samples useful controls?
  - Generate *in silico* samples instead

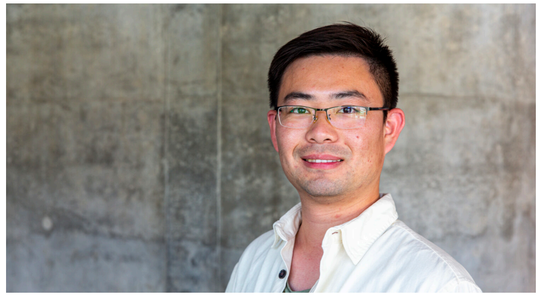
What is next?

- We wanna go to the real world and use our model
- Use cases:
  - Lets kill less mice
  - Rare diseases
  - And not also not so rare diseases

Acknowledgements



Acknowledgements



Acknowledgements