

PRECISION HEALTH  
ANALYSIS BOOTCAMP

# Supervised multiomics data integration

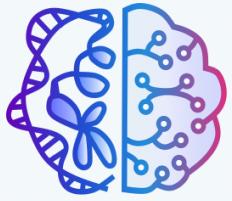
Amrit Singh, PhD

Department of Anesthesiology, Pharmacology and Therapeutics, UBC

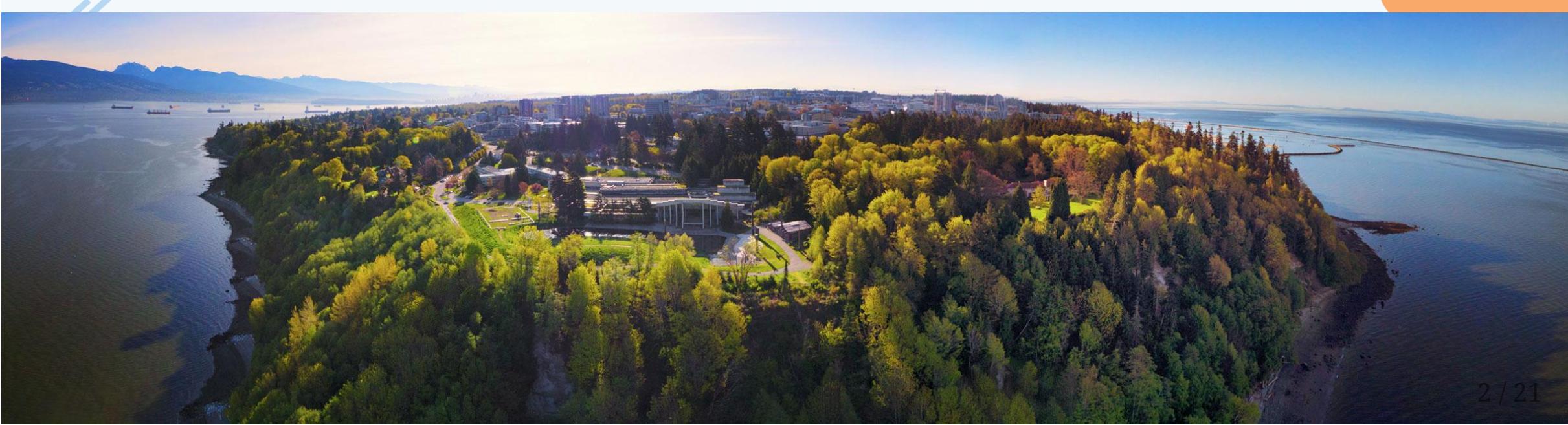
Centre for Heart Lung Innovation

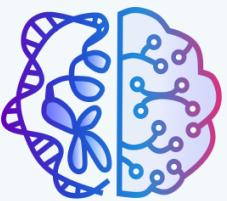
August 08, 2022 | 12:00-14:00





We would like to begin by acknowledging that the land on which we gather is the traditional, ancestral, and unceded territory of the xwməθkwəy̓əm (Musqueam) People.





# Copyright Information



## Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. [Disclaimer](#).

### You are free to:

**Share** — copy and redistribute the material in any medium or format

**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.



**Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



**ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

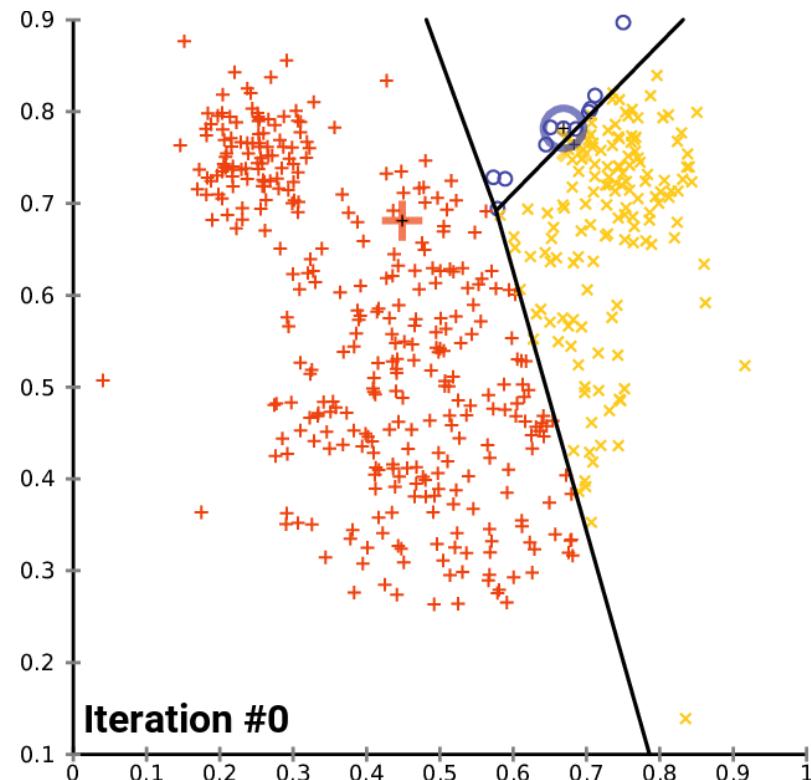
Read more here:  
<https://creativecommons.org/licenses/by-sa/4.0/>

# Learning outcomes

1. Contrast between unsupervised and supervised learning
2. Apply supervised integrative methods to real world data
3. Describe how biological information can be incorporated into supervised methods

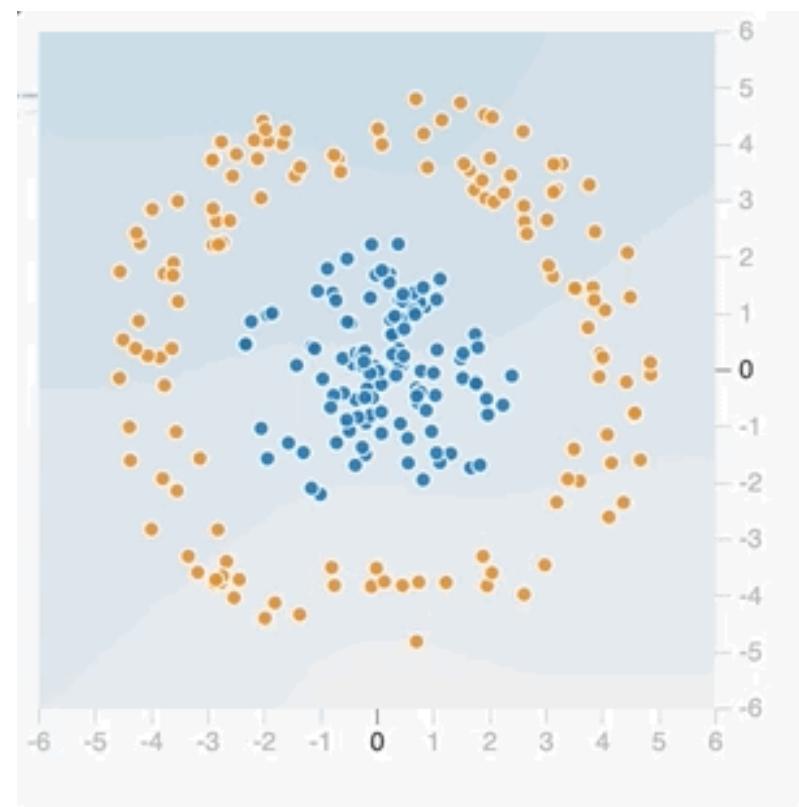
*"In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures."* Elements of statistical learning, 2008 page 3

## Unsupervised (clustering)



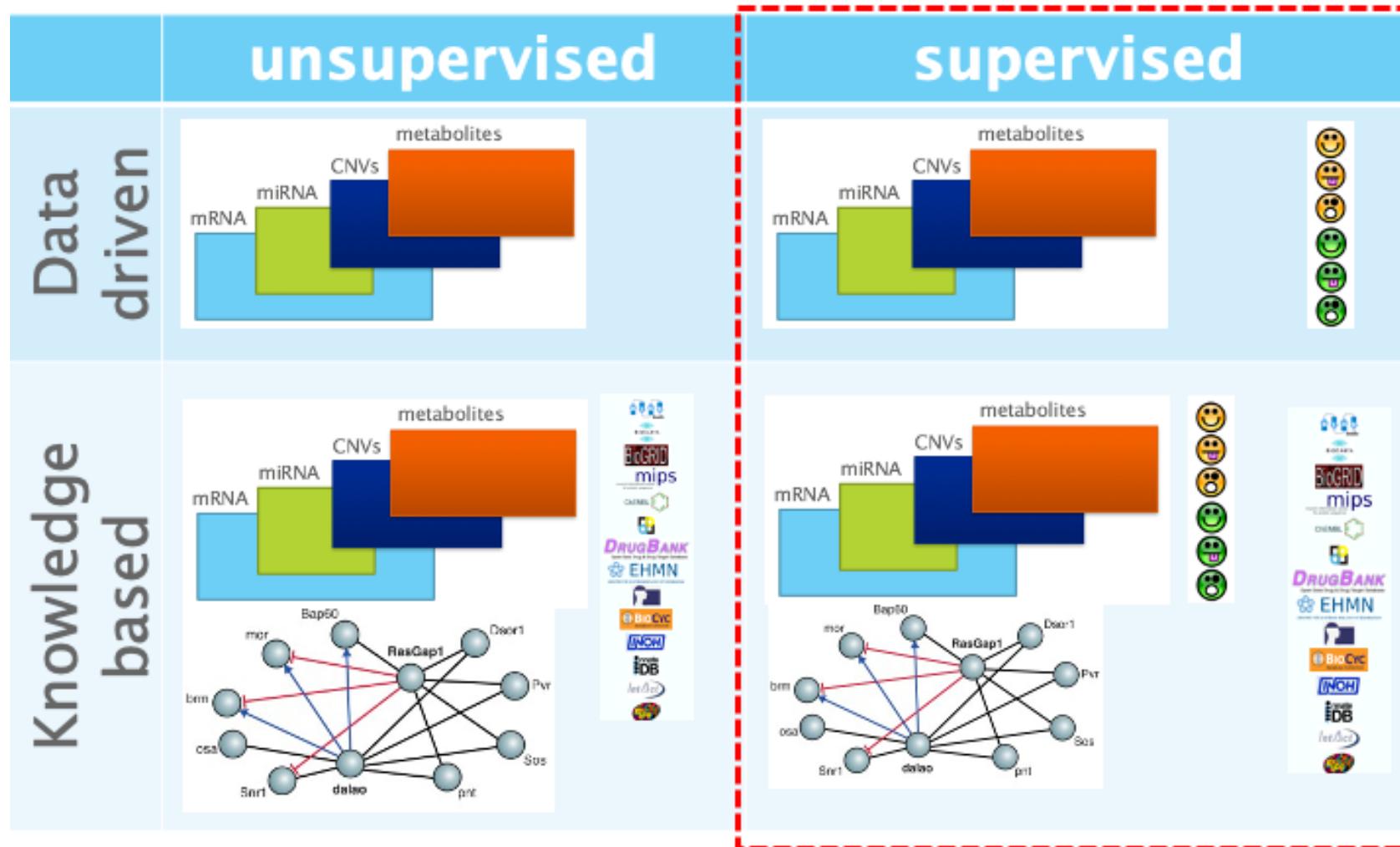
Chire 2017

## Supervised (regression/classification)

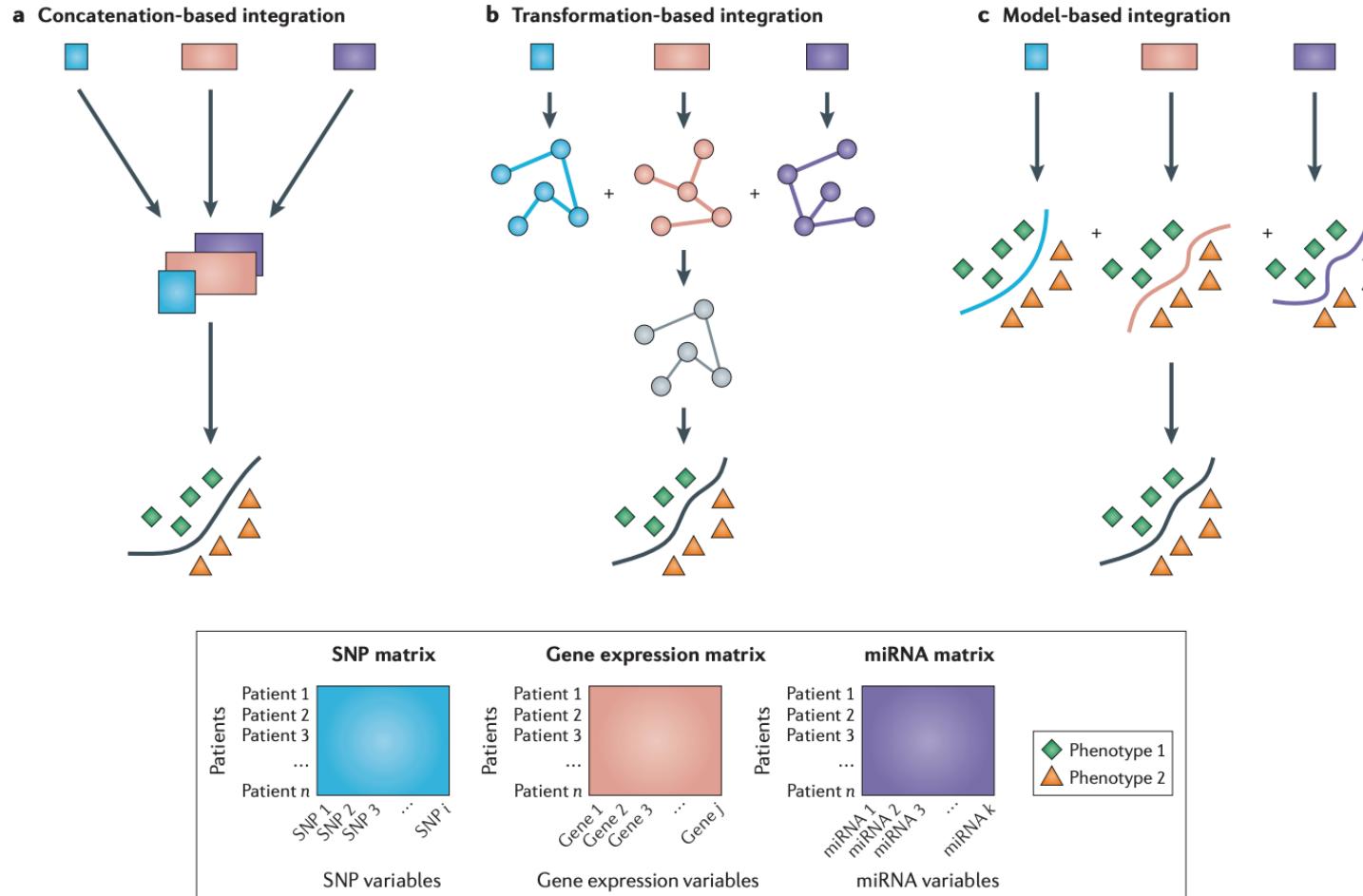


Tensorflow playground

# Multomics data integration



# Supervised multiomics data integration



# Dataset used in this talk

## Breast Cancer multi omics data from TCGA

This data set is a small subset of the full data set from The Cancer Genome Atlas that can be analysed with the DIABLO framework. It contains the expression or abundance of three matching omics data sets: mRNA, miRNA and proteomics for 150 breast cancer samples (Basal, Her2, Luminal A) in the training set, and 70 samples in the test set. The test set is missing the proteomics data set.

### Usage

```
library(mixOmics)
data(breast.TCGA)
```

# Concatenation-based

- strongest dataset will influence the biomarkers selected

```
concat_model <- splsda(X = do.call(cbind, breast.TCGA$  
                           Y = breast.TCGA$data.train$subty  
biomarkers <- selectVar(concat_model)$name  
  
p <- data.frame(dataset = names(breast.TCGA$data.train  
                           num = sapply(breast.TCGA$data.train[1:3], f  
ggplot(., aes(x = dataset, y = num)) +  
geom_bar(stat="identity")
```

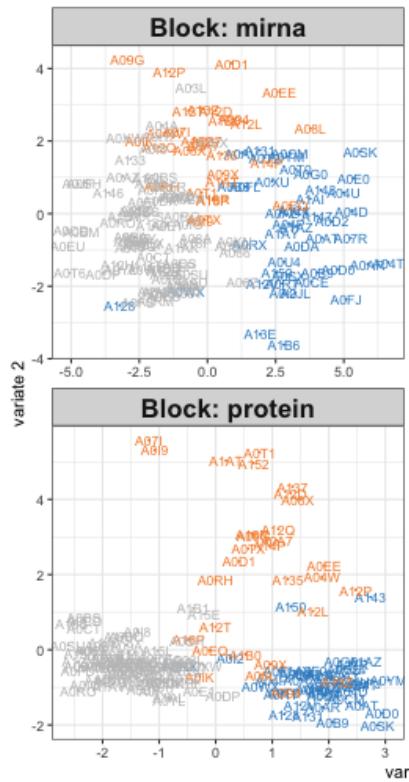
# Data Integration Analysis for Biomarker discovery using Latent c0mponents (DIABLO)

```
list.keepX <- list(mirna = c(16, 17), mrna = c(18,5),  
diabolo <- block.splsda(X = breast.TCGA$data.train[1:3]  
                           Y = breast.TCGA$data.train$subt  
                           keepX=list.keepX)
```

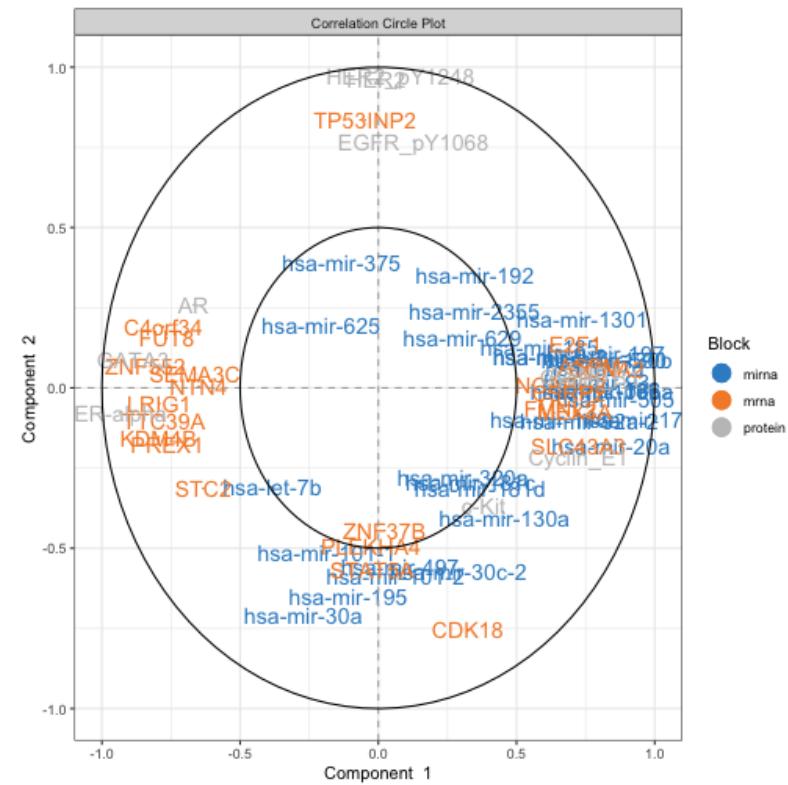
A horizontal progress bar consisting of a grey oval on a white background, indicating the status of the code execution.

# DIABLO

```
plotIndiv(diablo)
```



```
plotVar(diablo, var.names = c(TRUE, TRUE, TRUE),  
        legend=TRUE, pch=c(16,16,1))
```



# DIABLO

```
cimDiablo(diablo, color.blocks = c('darkorchid', 'brown1',  
comp = 1, margin=c(8,20), legend.position =
```

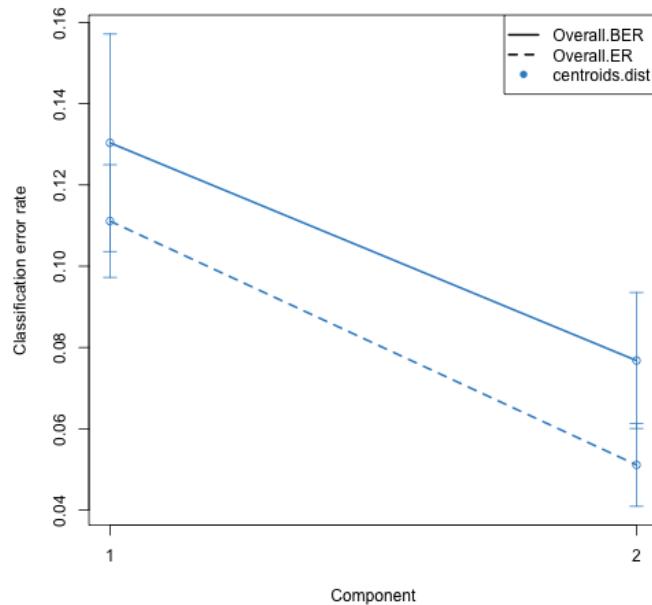
```
network(diablo, blocks = c(1,2,3),  
color.node = c('darkorchid', 'brown1', 'lightgreen1'),  
cutoff = 0.6)
```

```
##  
## trimming values to [-3, 3] range for cim visualisation. See 'trim' arg in ?cimDiablo
```

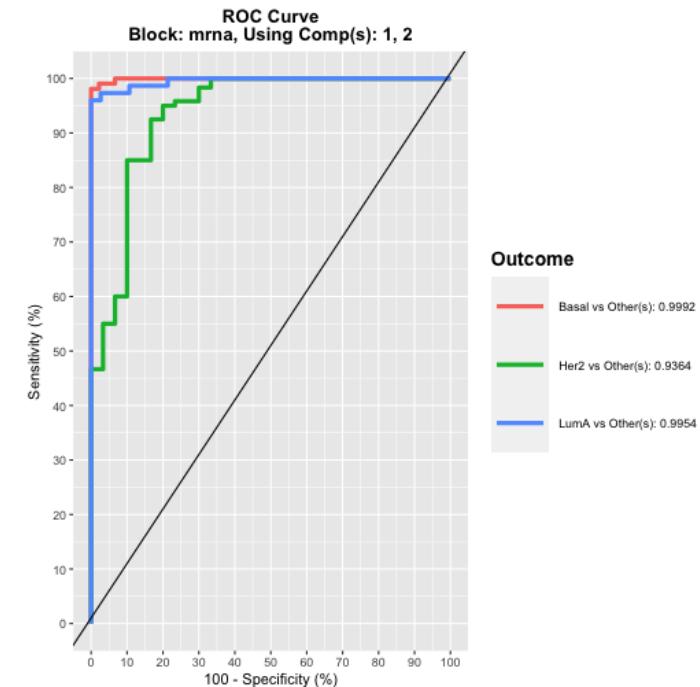
# DIABLO

```
set.seed(123) # for reproducibility in this vignette
perf_diabolo <- perf(diabolo, validation = 'Mfold', folds = 5,
                      nrepeat = 3,
                      dist = 'centroids.dist')
```

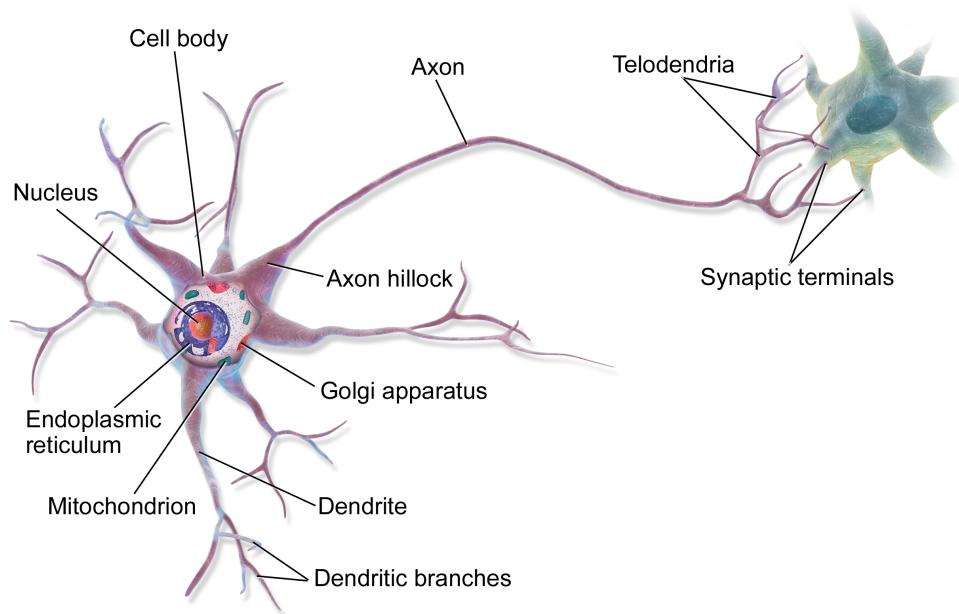
```
plot(perf_diabolo)
```



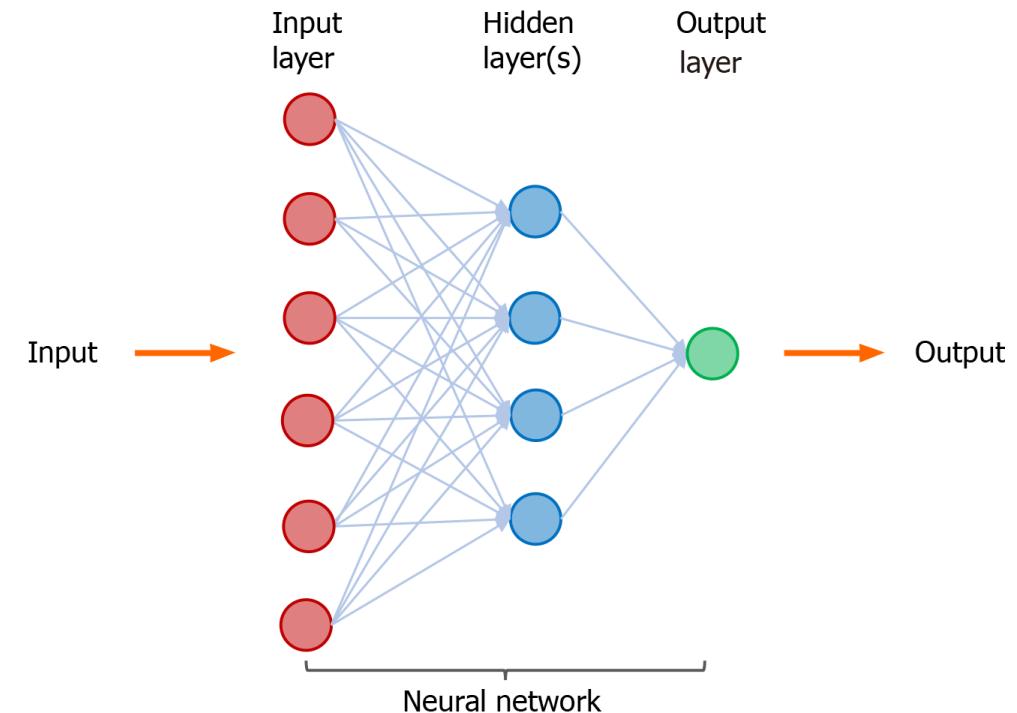
```
auroc(diabolo, roc.block = "mrna", roc.comp = 2)
```



# From biological to artificial neural networks (ANNs)

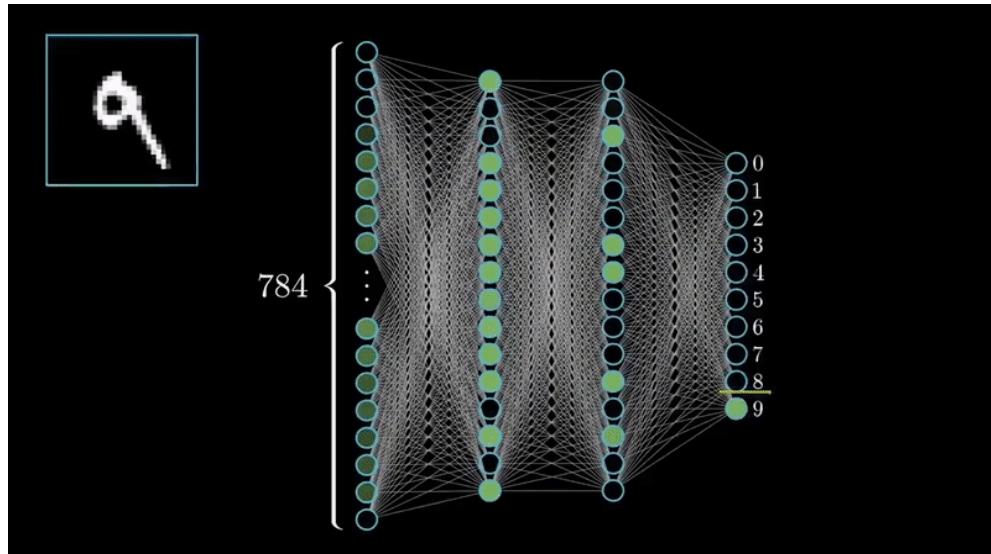


Neural circuit



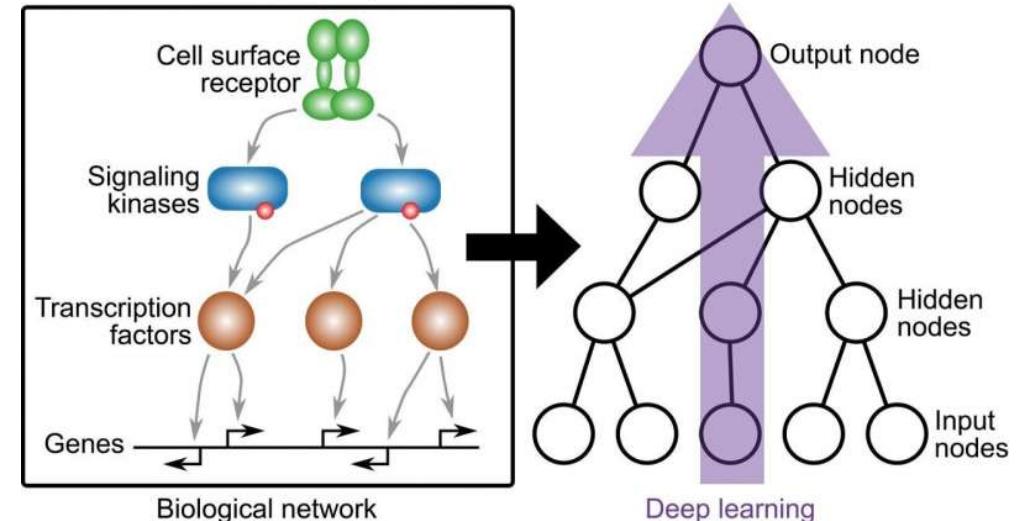
Yee NS. World J Transl Med 2021(9):1-10

# ANN structure captures hierarchical relationships



3brown1blue

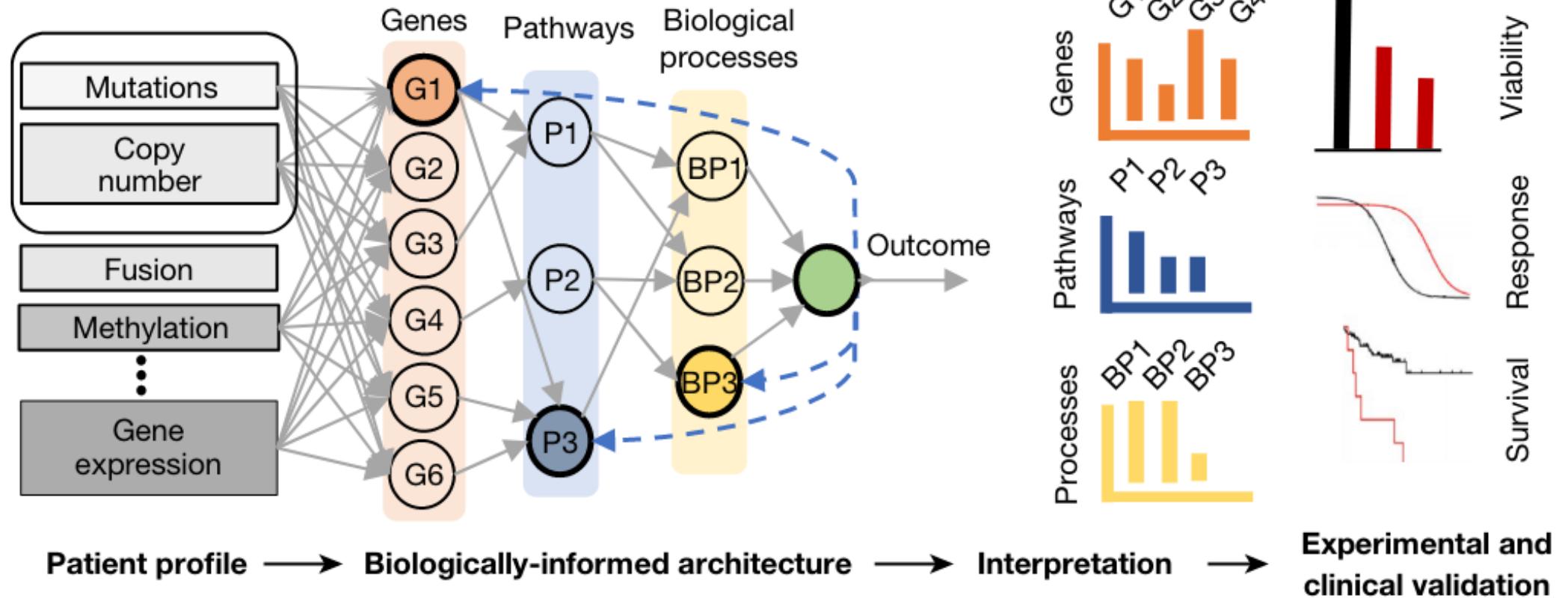
## Knowledge-primed neural networks



Knowledge-primed neural networks

Genome Biology volume 21, Article number: 190  
(2020)

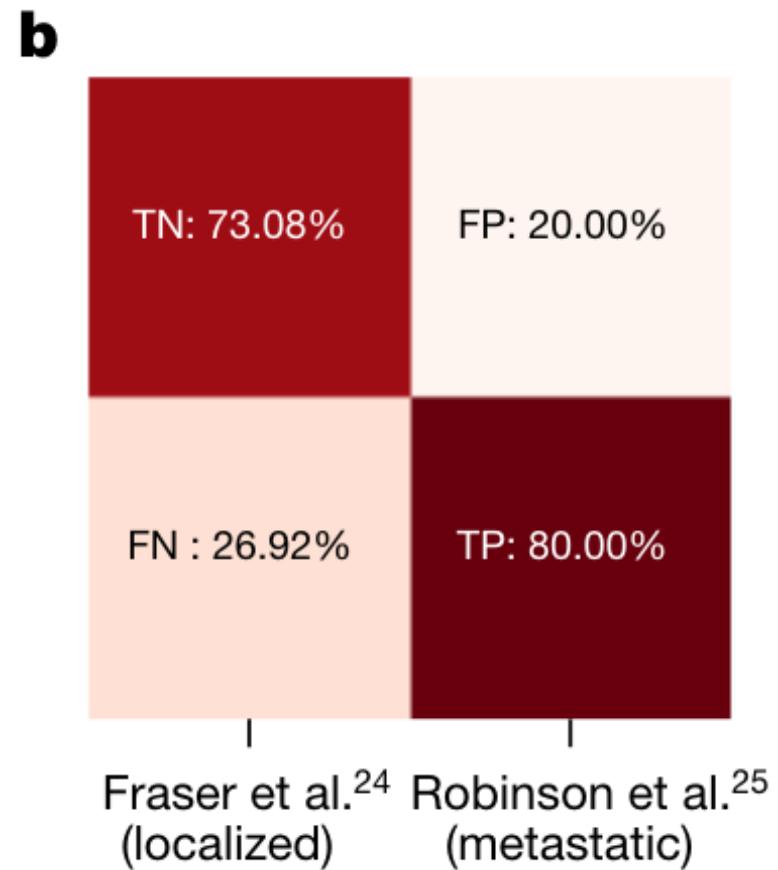
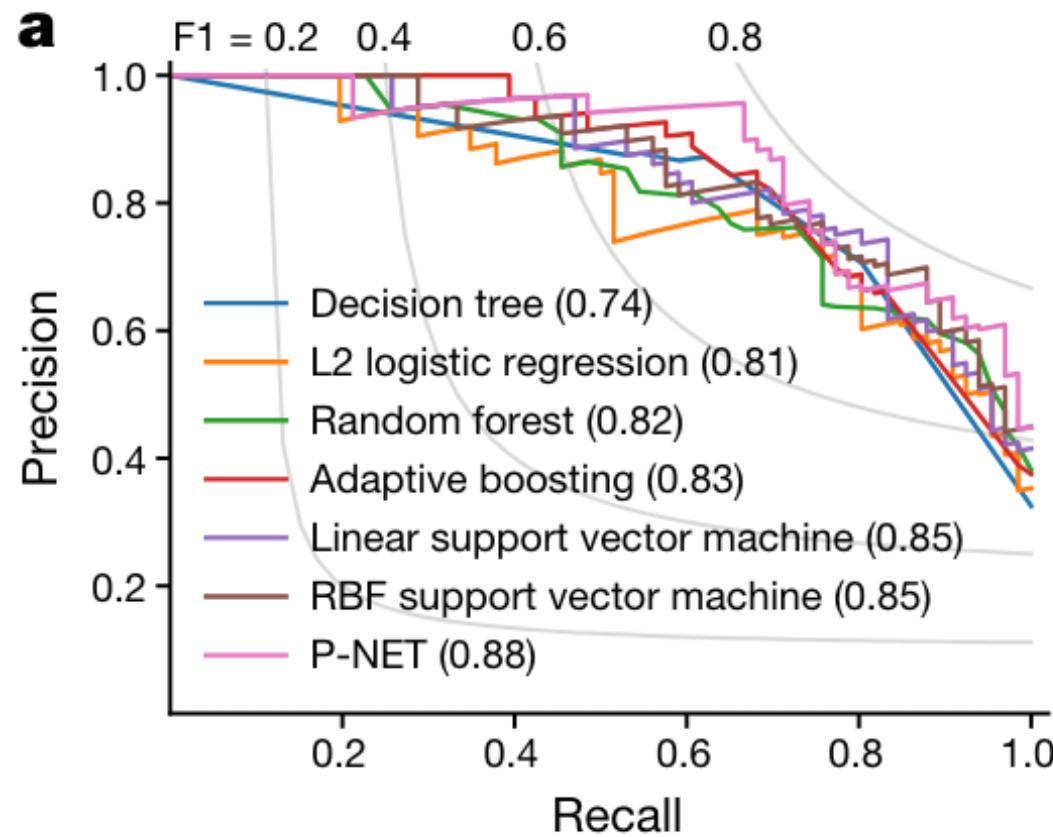
# Pathway-aware multi-layered hierarchical NETwork (P-NET)



Nature. 2021 Oct;598(7880):348-352.

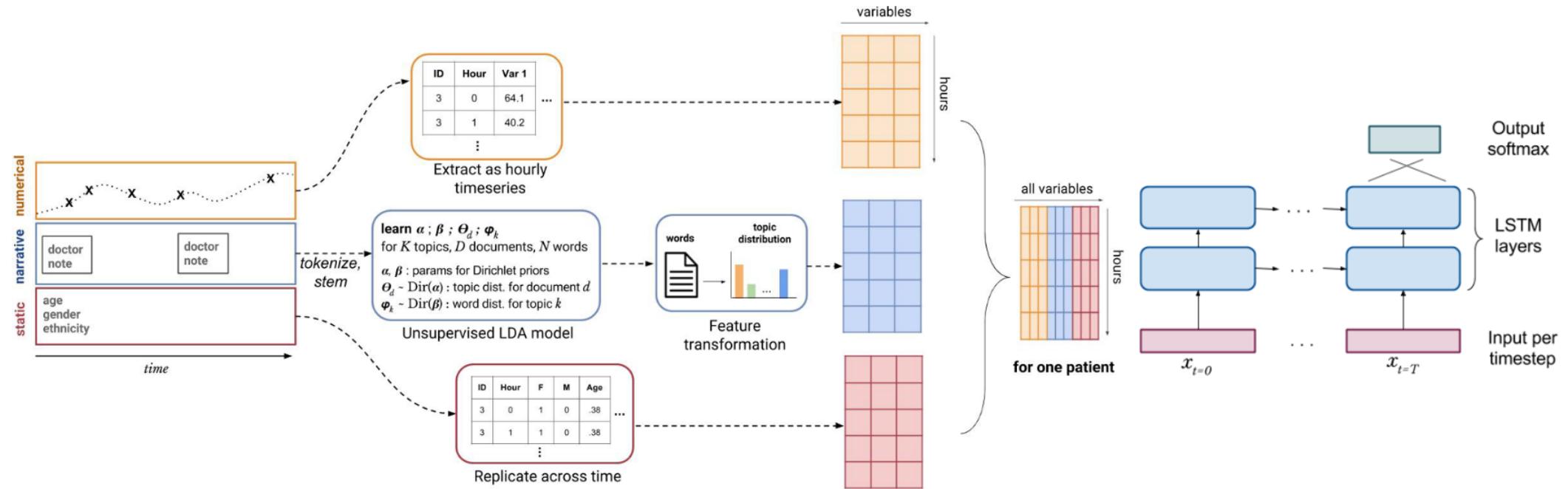
P-NET github

# P-NET



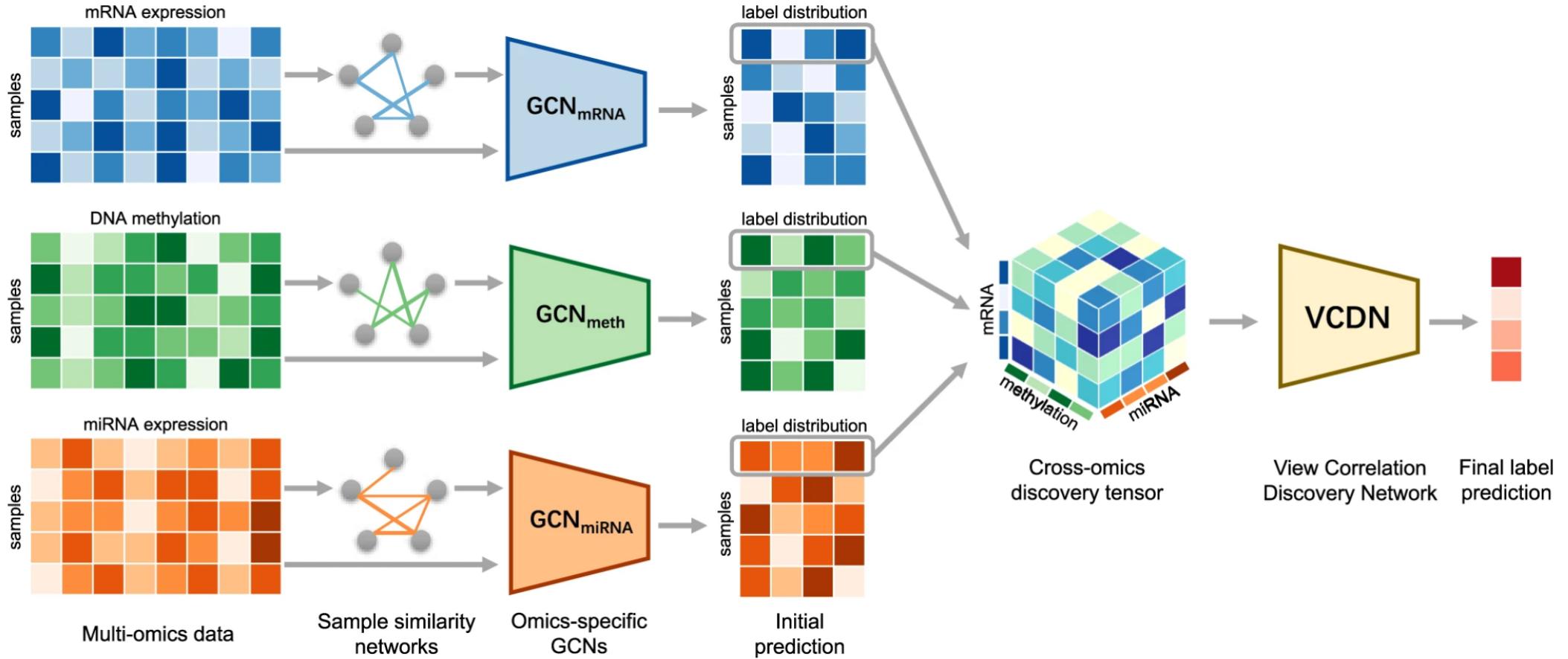
Nature. 2021 Oct;598(7880):348-352.

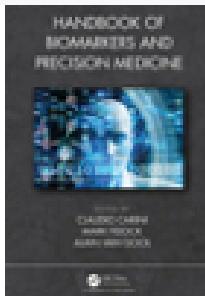
# Multimodal data integration



Suresh H et al., In Proceedings of Machine Learning for Healthcare 2017

# Multi-Omics Graph cOnvolutional NETworks (MOGONET)





## Unsupervised

## Supervised

MCIA: omicade4

sMB-PLS<sup>a</sup>

SGCCA: RGCCA/mixOmics

### Multiblock data analyses Extensions of generalized CCA

SMSMA: msma

SGCCA+DA: mixOmics

Data-driven

SNF:  
SNFtool

Message-  
passing  
algorithms

PANDA:  
pandaR

Joint NMP<sup>b</sup>

Factorization  
methods

SNMNMP<sup>c</sup>

BCC:  
bayesCC

Bayesian  
Methods

Bayesian  
Networks<sup>c</sup>

IBAG<sup>b</sup>

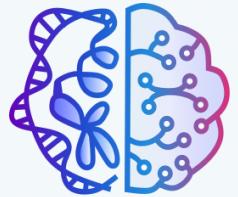
Concatenation: caret  
Ensemble: caretEnsemble

Classification and  
regression  
algorithms

stSVM: netclass  
GELnet: gelnet

Knowledge-  
based

Network-constrained



PRECISION HEALTH  
ANALYSIS BOOTCAMP

# THANK YOU!

August 08, 2022 | 12:00-14:00

lab  
code  
asingh\_22g