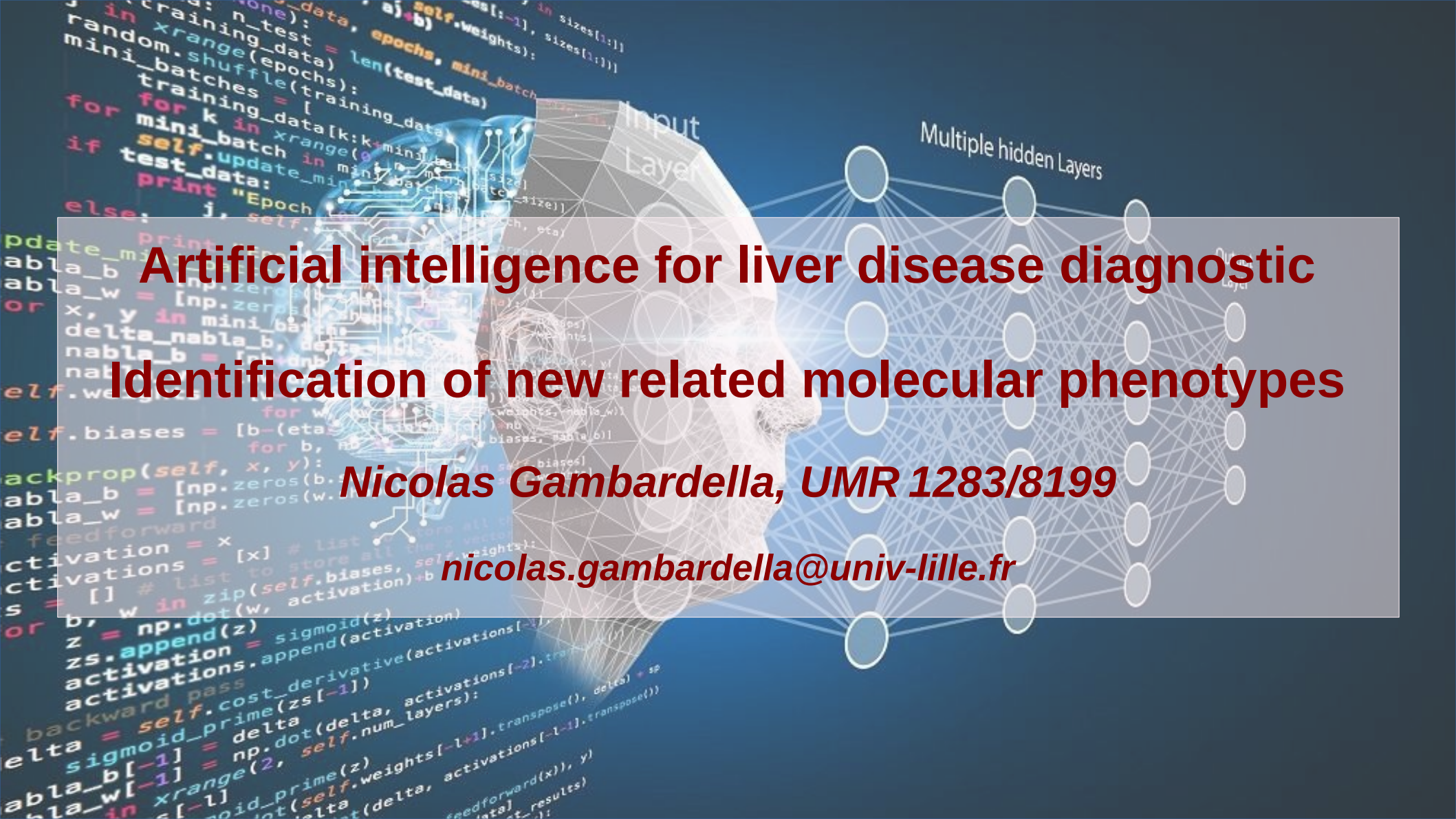


Artificial intelligence for liver disease diagnostic

Identification of new related molecular phenotypes

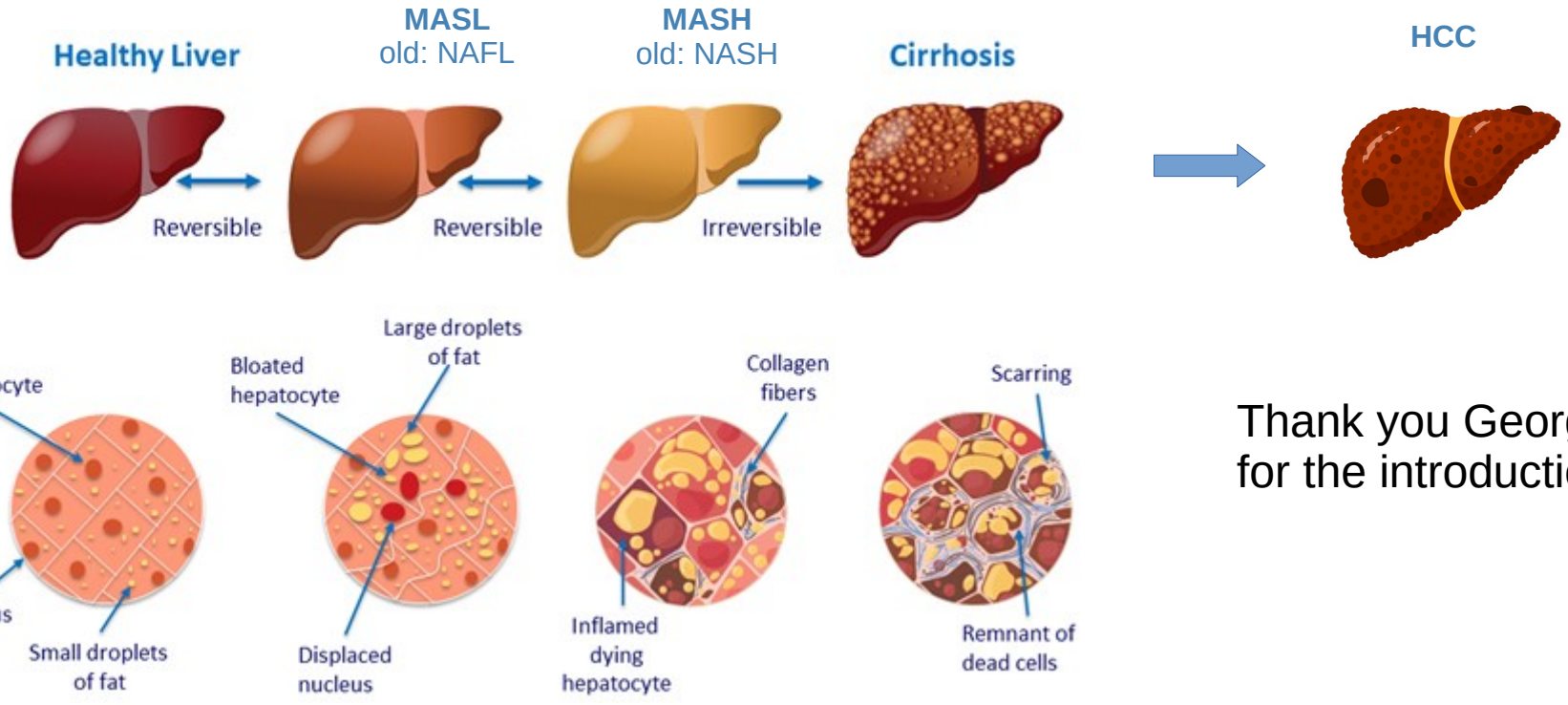
Nicolas Gambardella, UMR 1283/8199

nicolas.gambardella@univ-lille.fr

The spectrum of non-alcoholic steatotic liver disease (SLD)

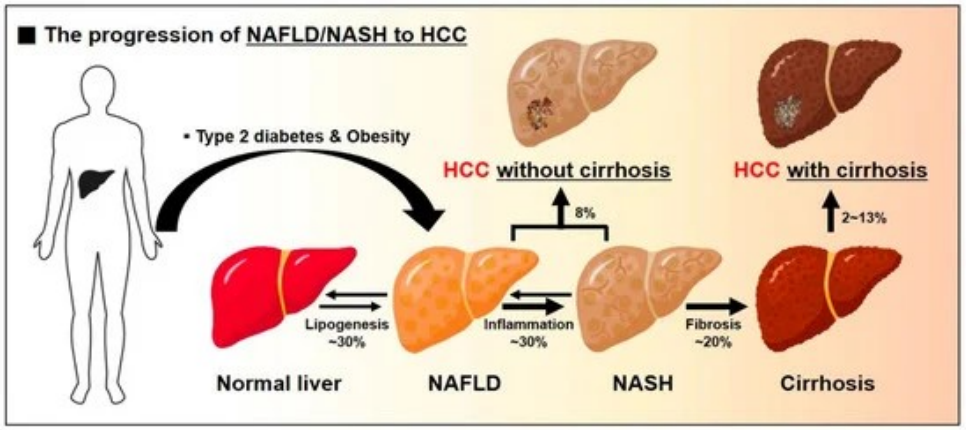
stéar (στέαρ) = fat

Metabolic dysfunction-associated steatotic liver disease (MASLD)

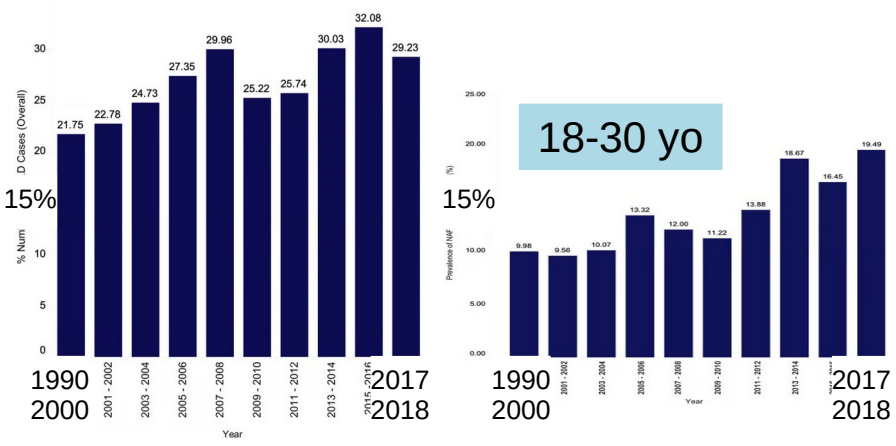


Thank you Georgia for the introduction

MASLD prevalence is growing

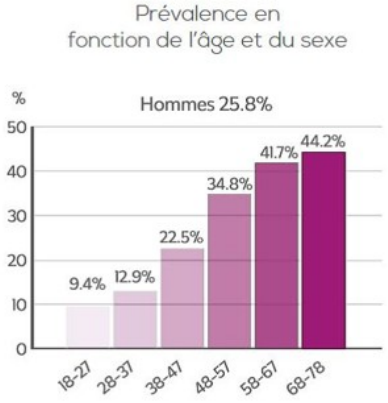


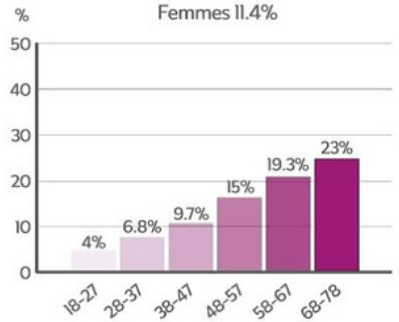
Kim *et al* (2021) *Int. J. Mol. Sci.*, 22(9): 4495

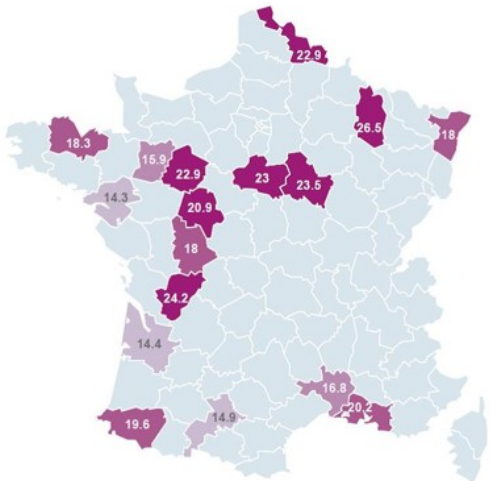


Kim *et al* (2022) *Met. Target Organ Damage*, 2: 19

MASLD





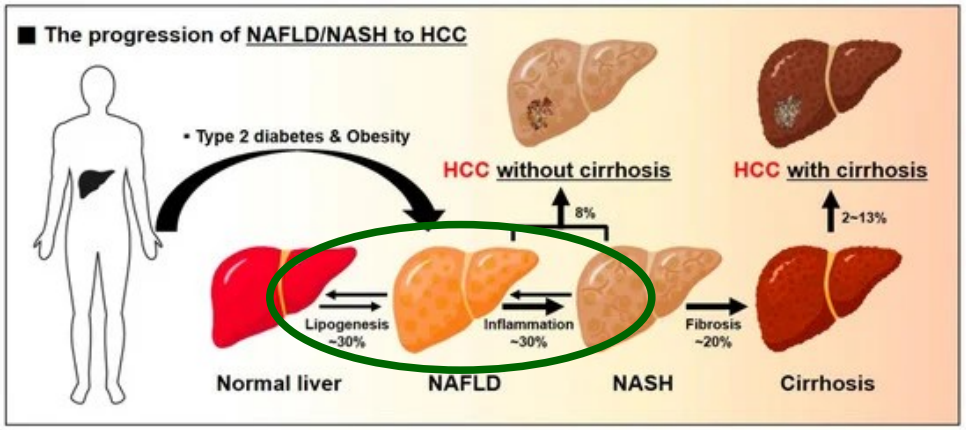


Répartition par régions

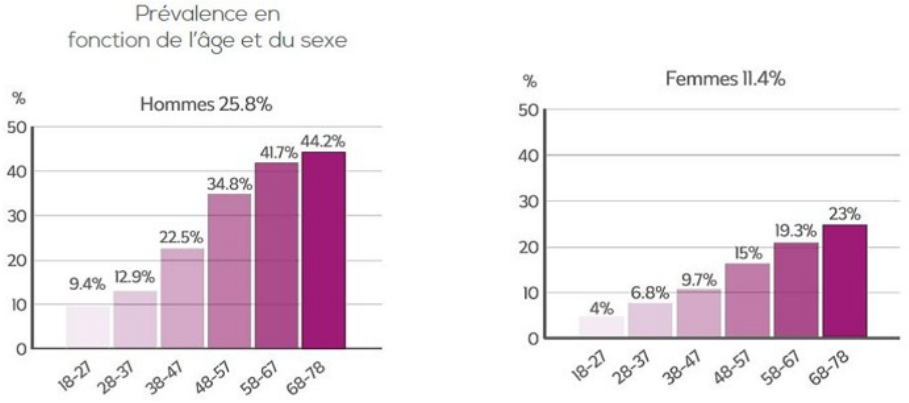
MASH

Paris Nash Meeting
(11-12 juillet 2019)

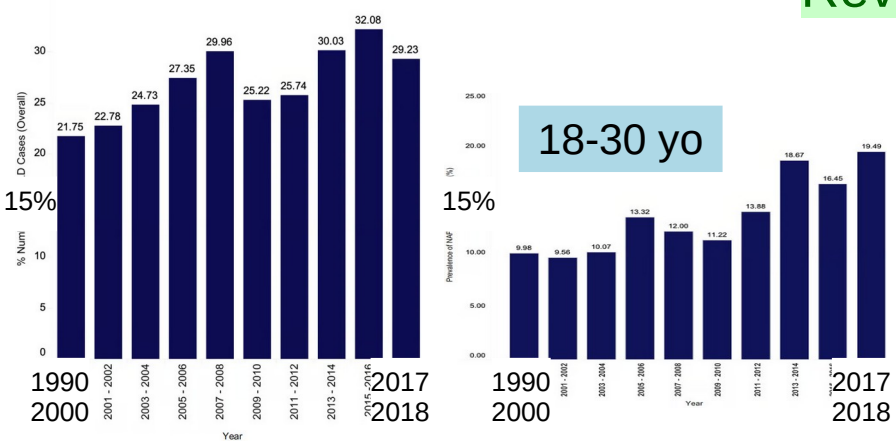
MASLD prevalence is growing



Kim *et al* (2021) *Int. J. Mol. Sci.*, 22(9): 4495



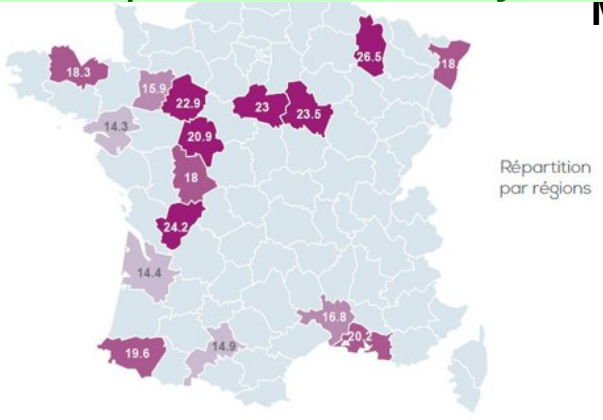
Reversibility = importance of early detection



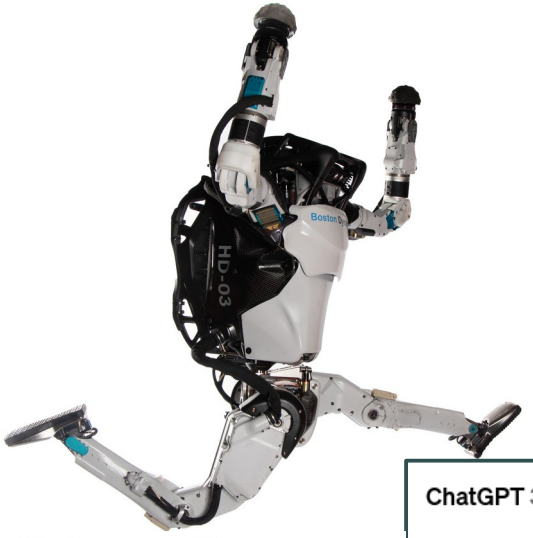
Kim *et al* (2022) *Met. Target Organ Damage*, 2: 19

MASLD

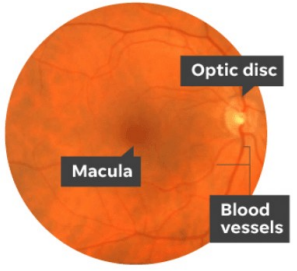
MASH

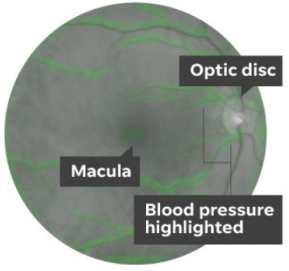


Paris Nash Meeting
(11-12 juillet 2019)



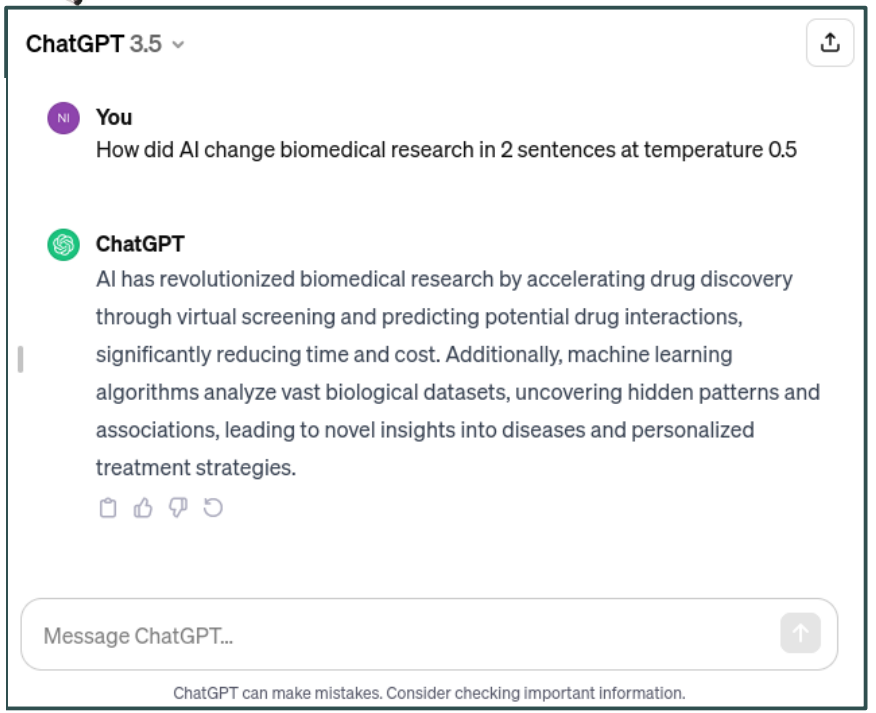
Google AI can predict heart problems by taking pictures of your eye





- Image of the back of the eye showing the macula (dark spot in the middle), optic disc (bright spot at the right), and blood vessels.
- Retinal image in gray, researchers can focus on blood vessels to determine the health risks associates with a patient's blood pressure.
- Images showed that each cardiovascular risk factor prediction uses a distinct pattern, such as blood vessels for blood pressure and optic disc for other predictions.

Source: Baig, Edward C. "Google Hopes AI Can Predict Heart Disease by Looking at Retinas." *USA Today*, Gannett Satellite Information Network, 19 Feb. 2018, www.usatoday.com/story/tech/2018/02/19/google-ai-can-predict-heart-disease-looking-pictures-retina/344547002/.

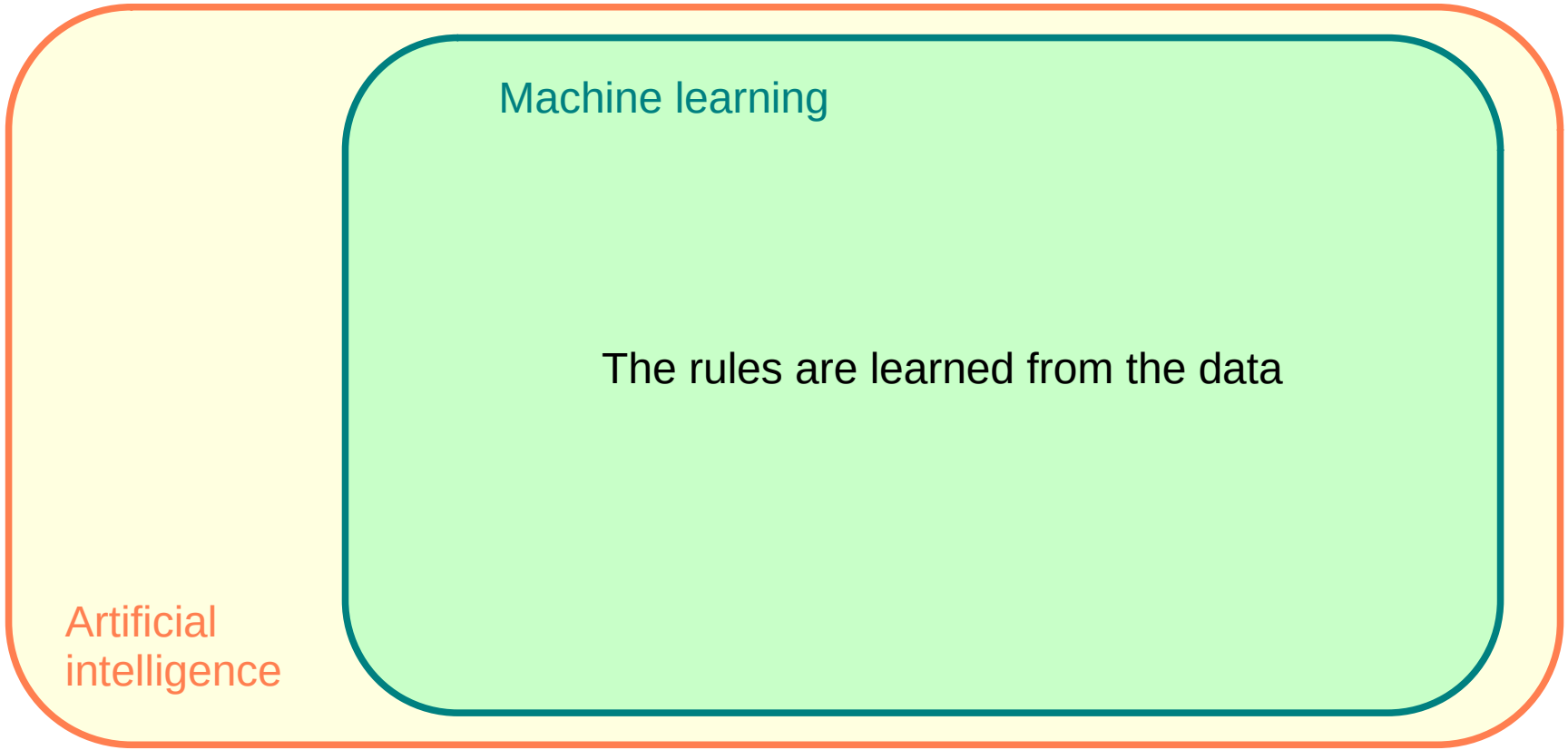


Some terminology

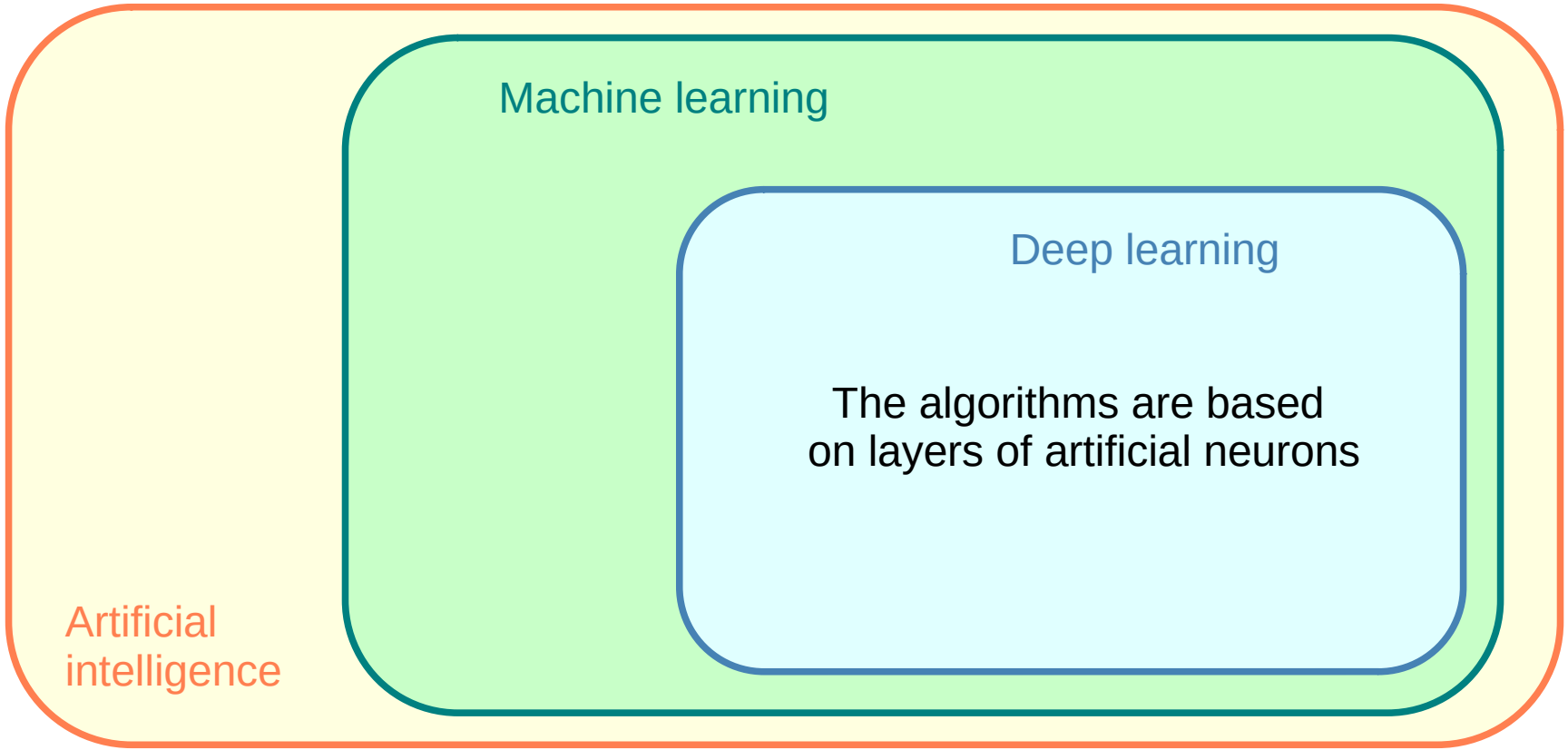
Assessments, evaluations, decisions, predictions
made by software tools

Artificial
intelligence

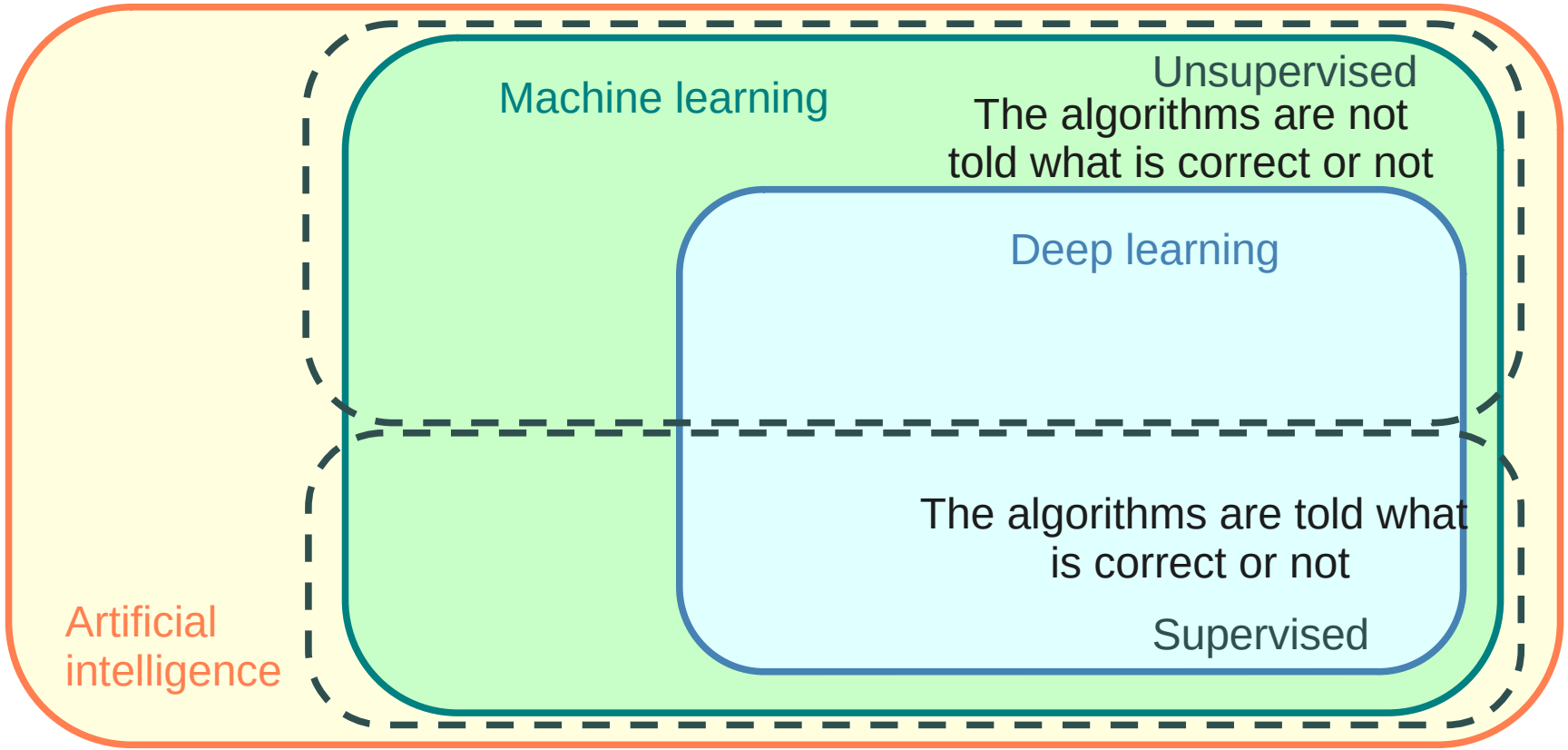
Some terminology



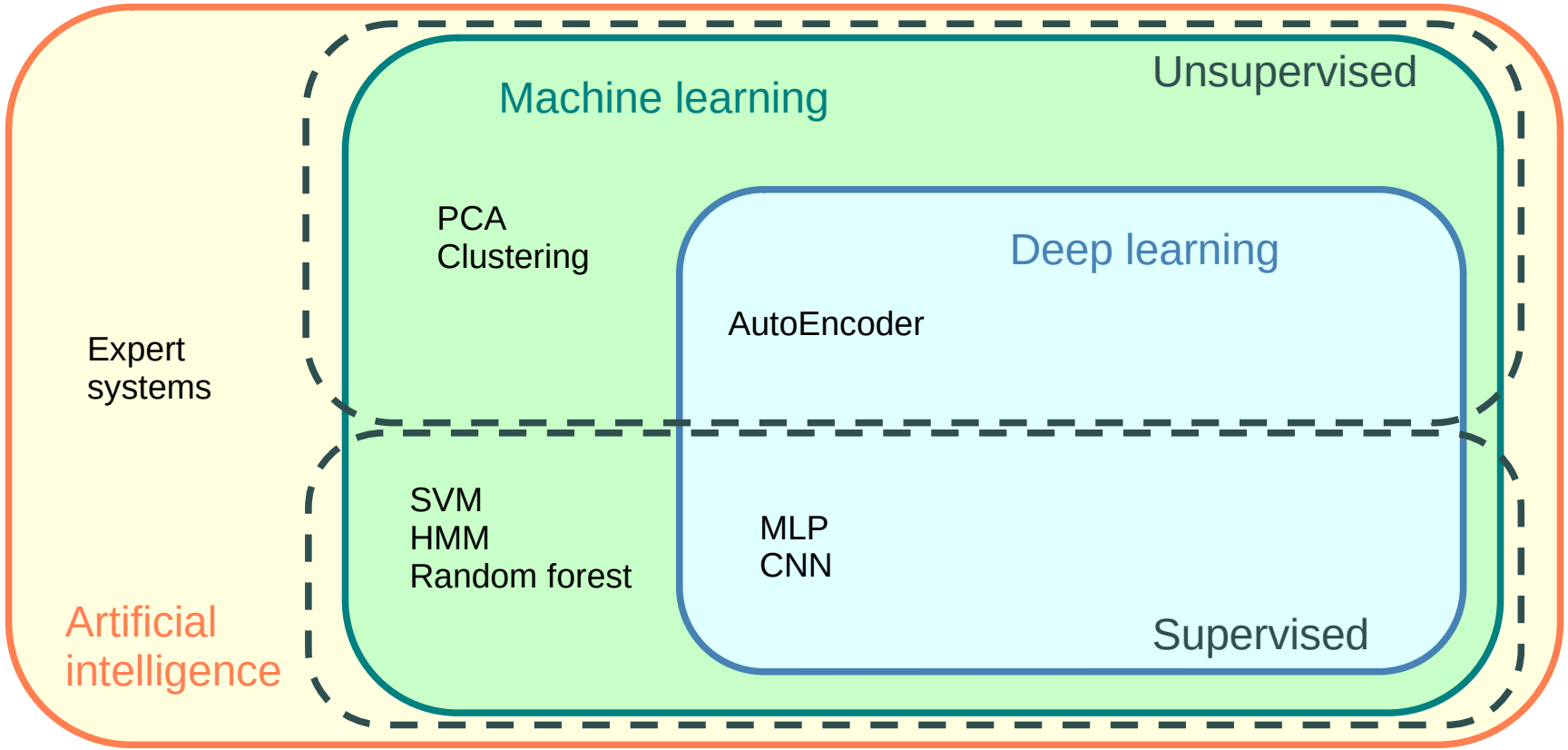
Some terminology



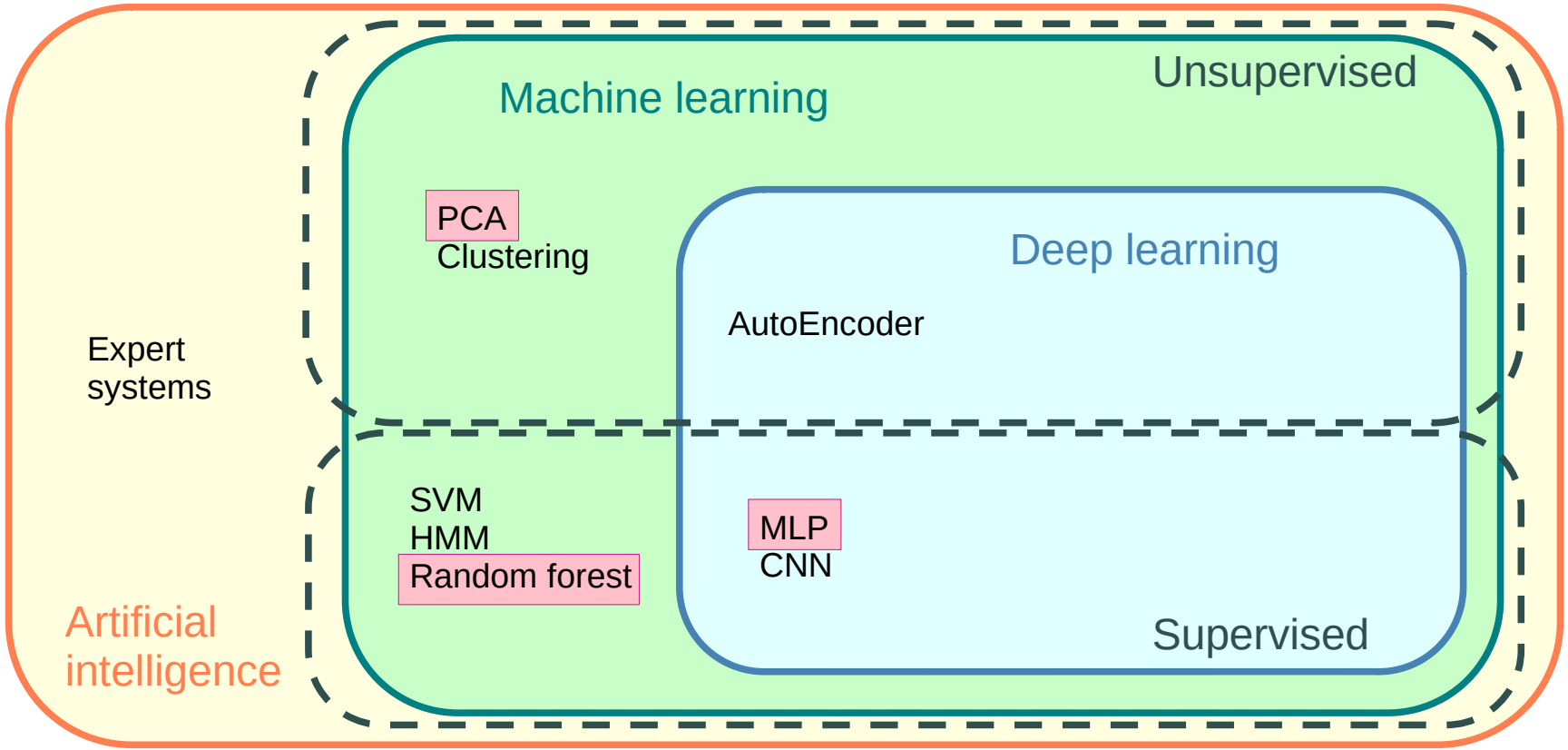
Some terminology



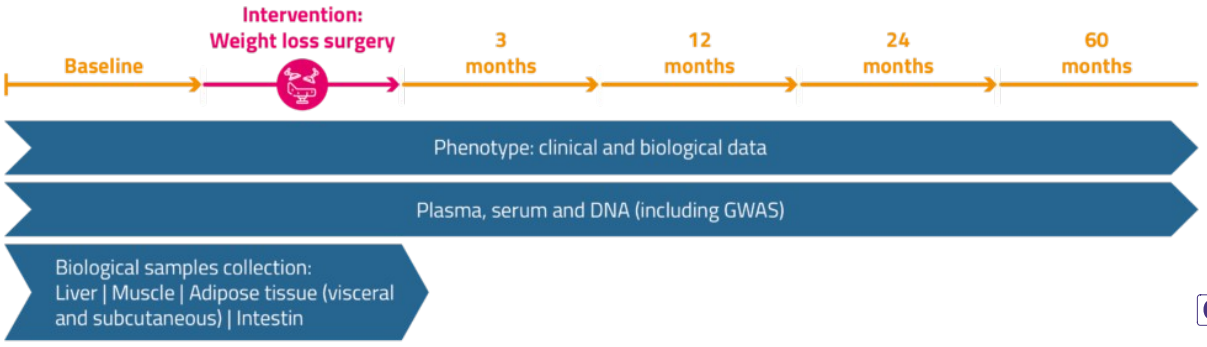
Some terminology



Some terminology



Cohort

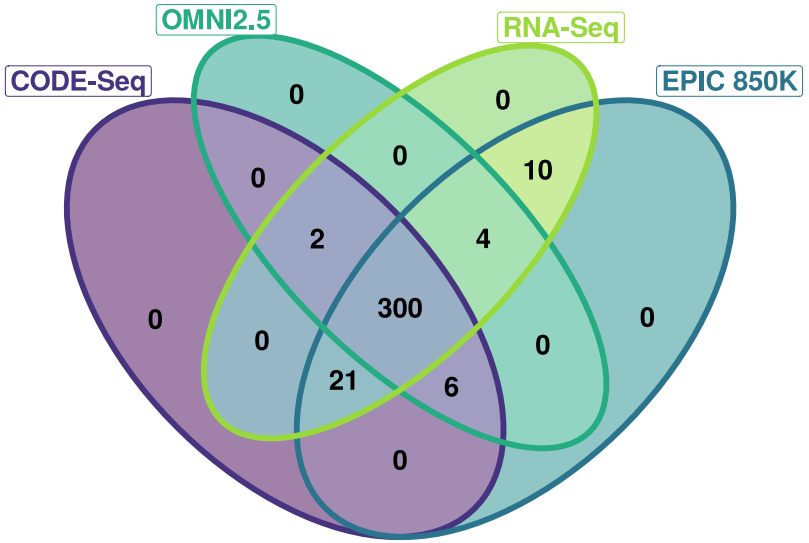


ABOS (Biological Atlas of Severe Obesity)

All subjects had bariatric surgery

PreciNASH project

ABOS subset: Only European ancestry and unrelated individuals



(+66 clinical and personal data
+1076 identified metabolites
in blood and liver)

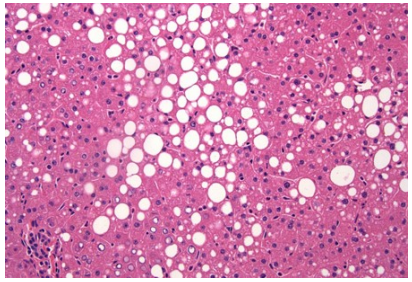
Subject grouping

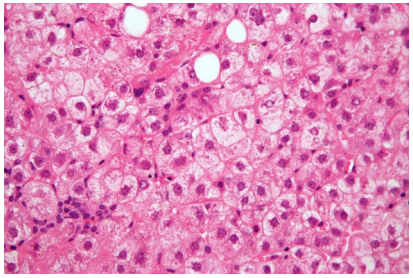
Scoring on liver biopsy with the method from Kleiner and Brunt 2005

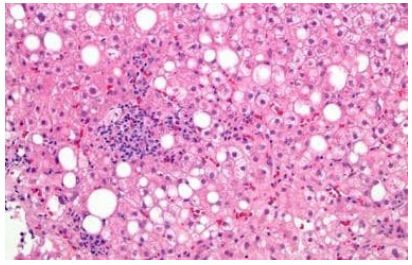
Steatosis
Categorical [0-3] from quantitative measurement

Ballooning
Categorical [0-2]
= {none, some, much}

Inflammation
Categorical [0-3] from number of foci







Final score:

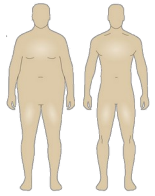
Healthy: S = 0, B = 0, I = 0 n = 80

NAFL: S > 1, B = 0, I ≥ 1 n = 137
S > 1, B > 1, I = 0

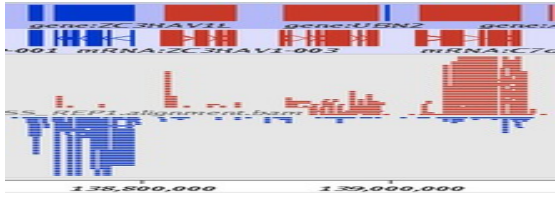
NASH: S > 0, B > 0, I > 0 n = 83

What are we going to talk about today?

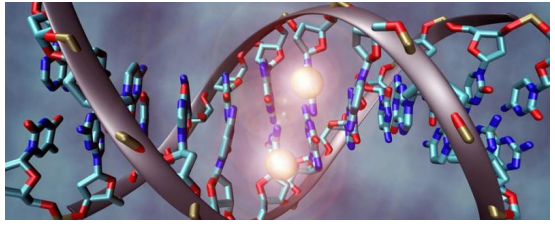
- Clinical data

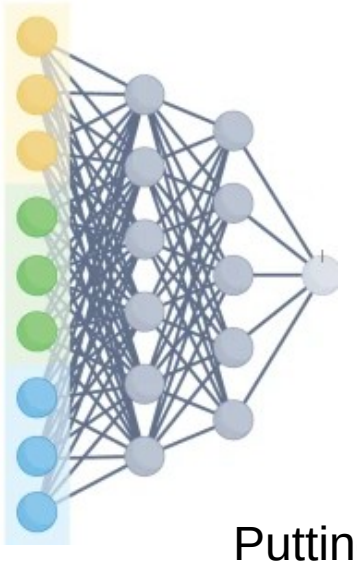


- RNAseq



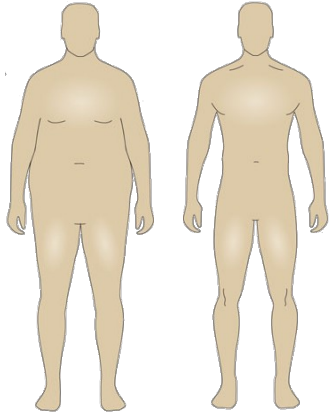
- CpG Methylation

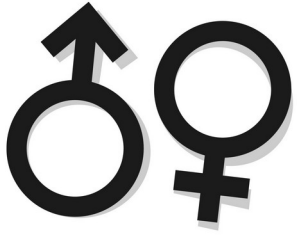




Putting everything together

Clinical data





Clinical data available on PreciNASH subjects

Diagnosis
Sex
Age
BMI

Liver

Steatosis score
Inflammation score
Ballooning score
Brunt score
Total NAS score
Fibrose Kleiner score

Sang

Glycaemia: fasting ionograms,
OGTT: fasting, 30 min, 120 min
Insulin: fasting (mUI/L, pmol/L), 30 min, 120 min
C peptide: fasting, 30 min, 120 min
HOMA2 IR
HbA1c
HDL
LDL
Total cholesterol
Triglycerides
Bilirubin
ASAT
ALAT
γGT
α2 macroglobulin
Haptoglobin
apoA1
CRP
platelets
lymphocytes

Hypertension

Antihypertensive (name)
α–blocker
β–blocker
Ca blocker
Angiotensin inhibitor
Imidazoline
ACE inhibitors
Diuretics
Renin inhibitor
α–sympathomimetic
α1-adrenergic blocker

Diabetes

T2D status
antidiabetic treatments (number and name)
Insulin treatment
GLP1 treatment
Gliptine
α-glucoxidase inhibitor
Sulfonylureas
Biguanide
Glinide
Thiazolidinediones

Dislipidaemia

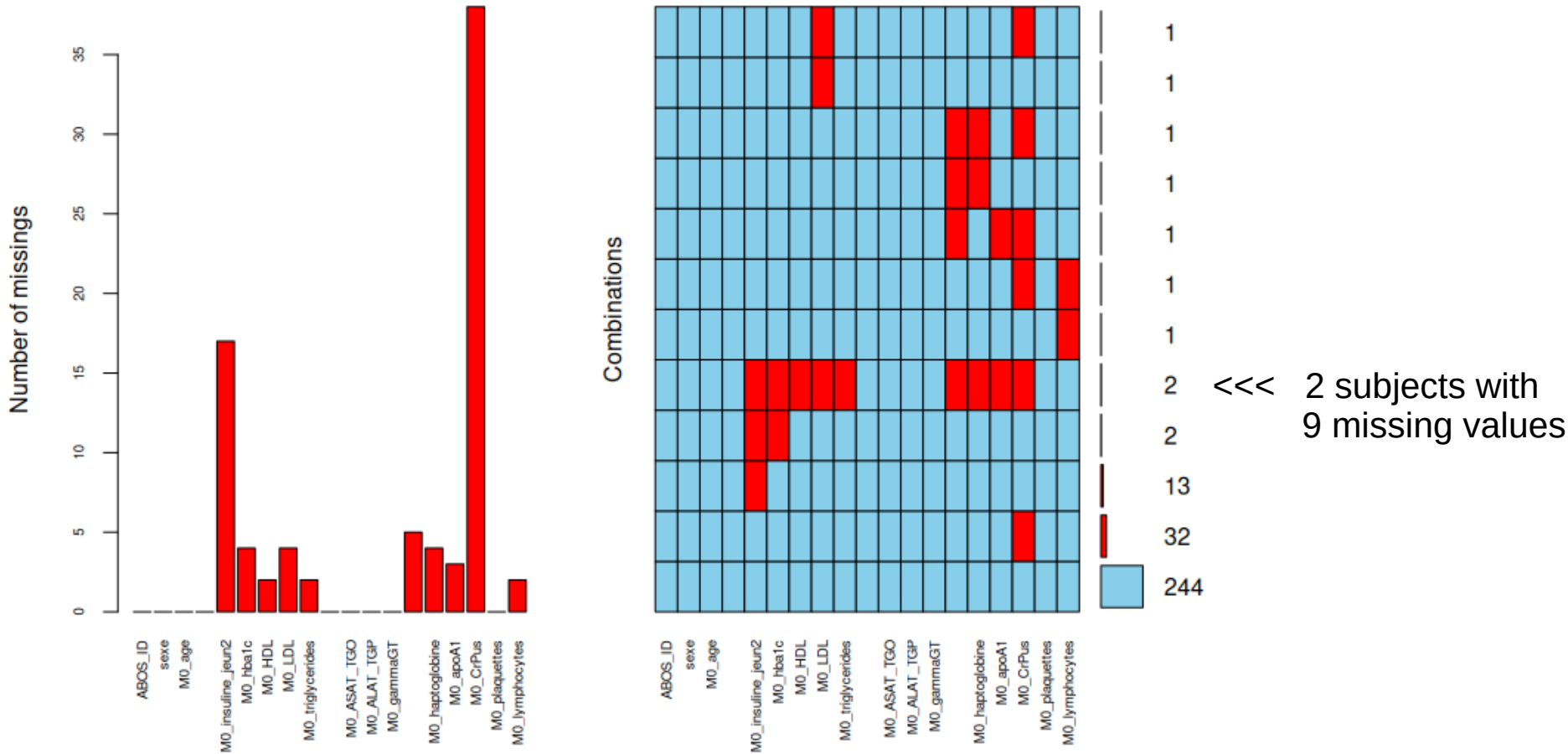
Hypolipemic treatment
(use and name)
Statins
Fibrates
Omega-3
Intestinal absoption inhibitor
Bile acid sequestrant

Selection of clinical variables

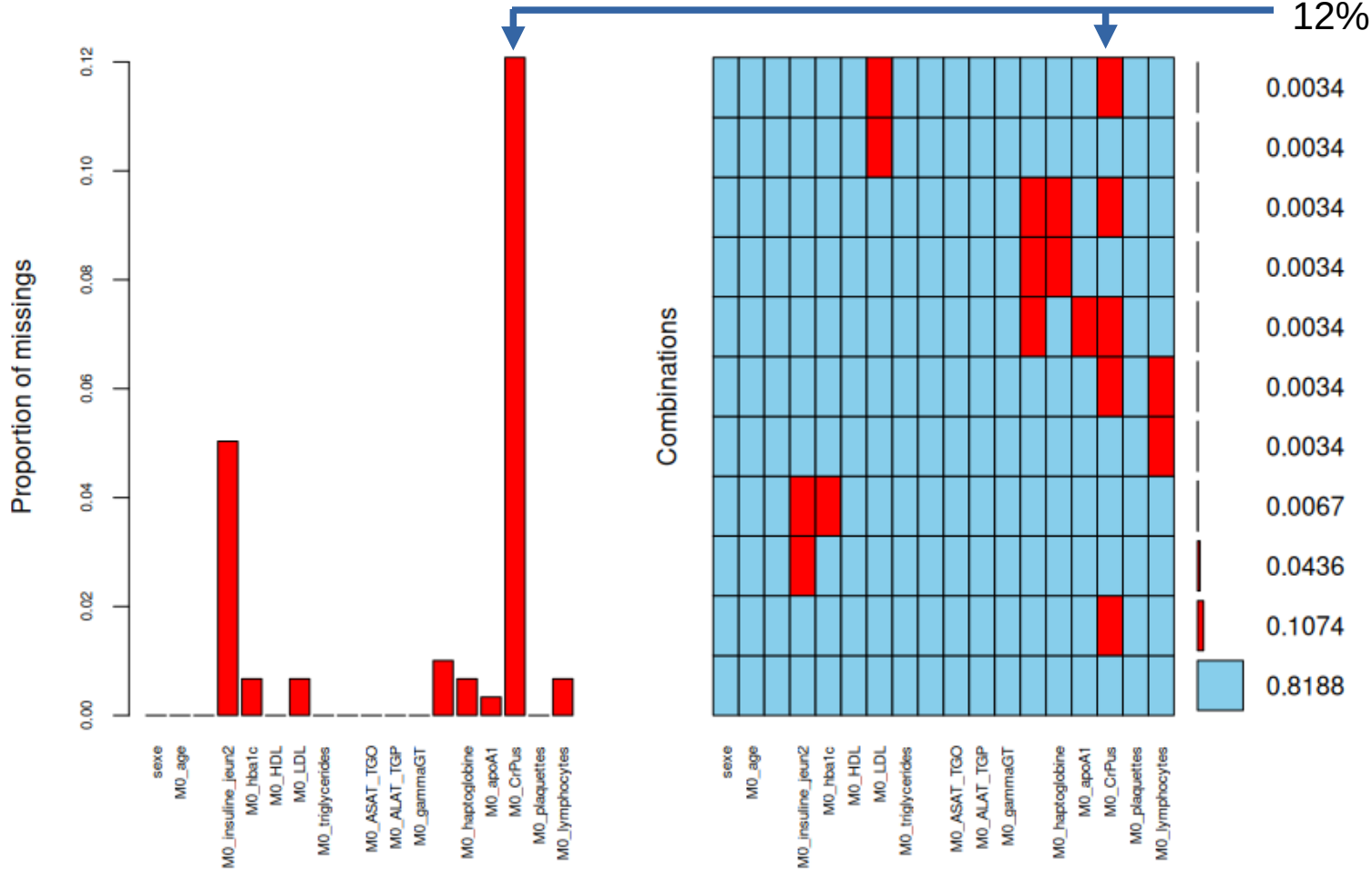
- Alternative versions: fasting insulin in mIU/L and in pm/L
- Composed variables: HDL, LDL, total cholesterol
- Derived variables:
  T2D status = {fast glyc, HbA1c, glycaemia, antidiabetic treatment}
  Averaged fasting glycaemia = mean(HPO glyc, IONO glyc)
  HOMA2 IR = {OGTT and insulin fasting}
- Treatments: ignored in this study that focuses on biology

☛ retained 16 non-redundant clinical features + age + sex

Clinical data: Missing values; small dataset = need imputation



Clinical data: Missing values; small dataset = need imputation



12% miss CRP us

Advised cut-off for imputation 5%

But CRP is key in MASLD…

Imputation of missing values: MICE

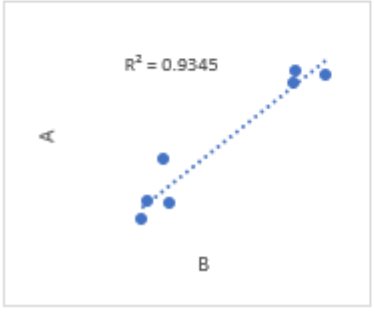
- General idea: subjects with close values for most variables would probably show similar values for the missing variables.
- Chosen approach is *Multivariate Imputation by Chained Equations* (R package *MICE*)
  Imputation algorithm is *Predictive mean matching*

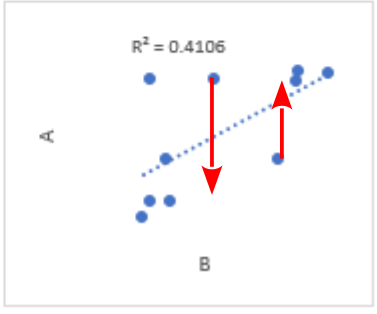
A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45

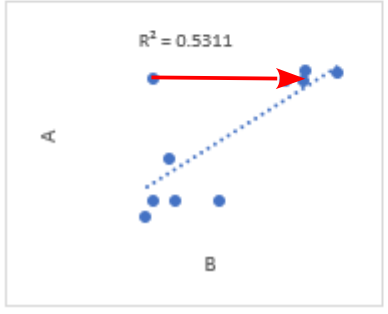
A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45

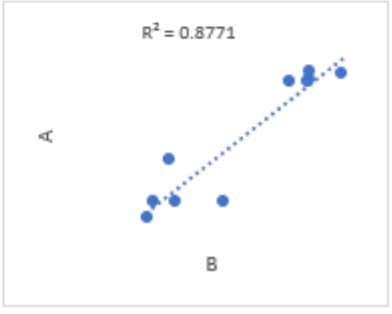
A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45







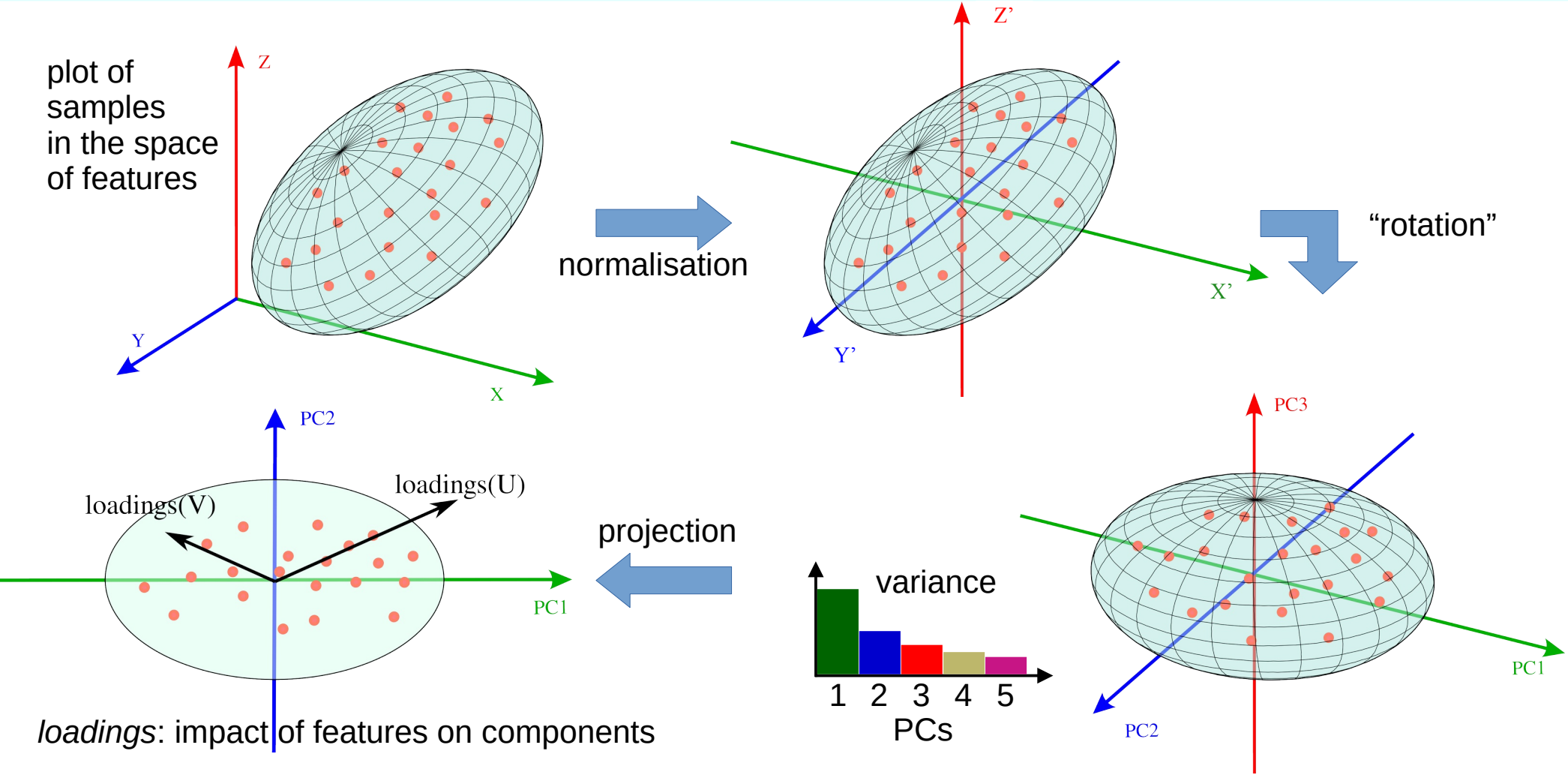


Then iteration

NB: no linearity or normality requirements

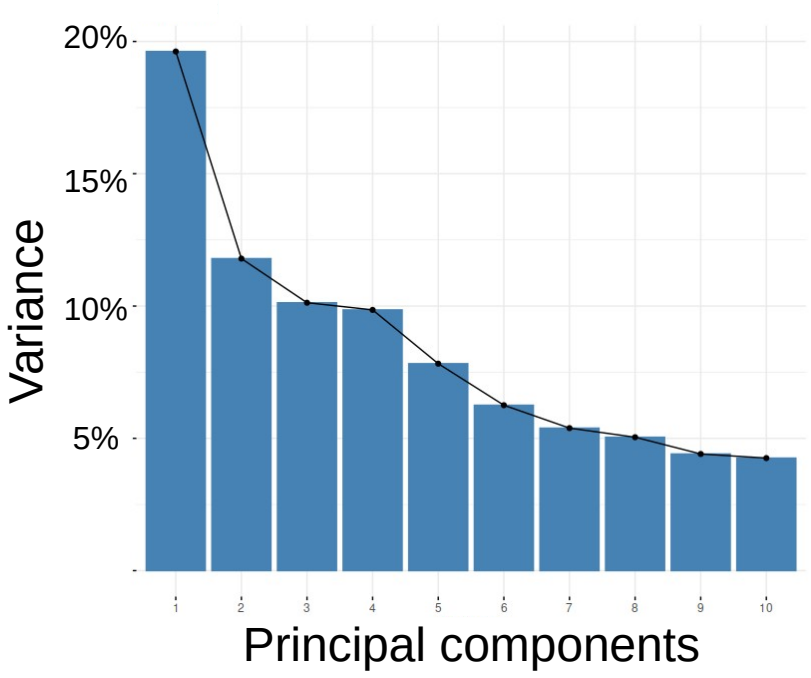
Suitable for multimodal, integer, or skewed variables

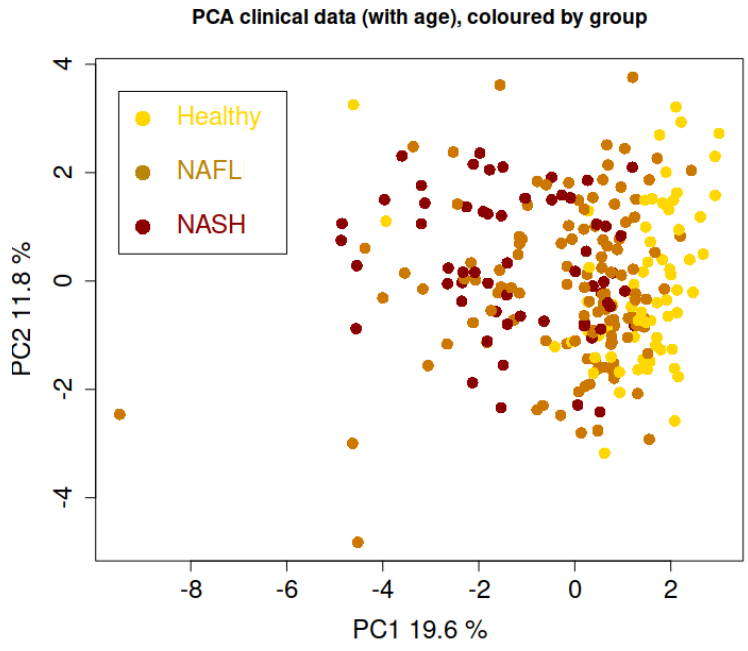
Principal component analysis (PCA)



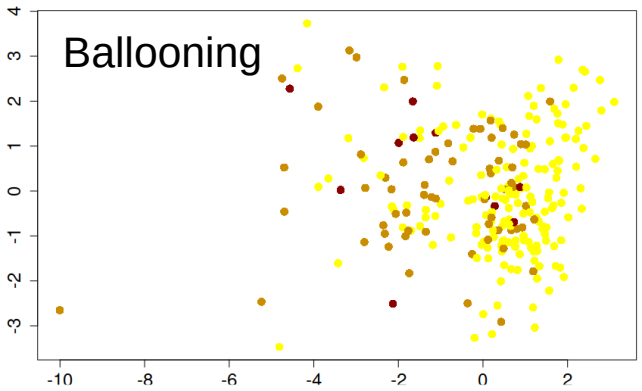
loadings: impact of features on components

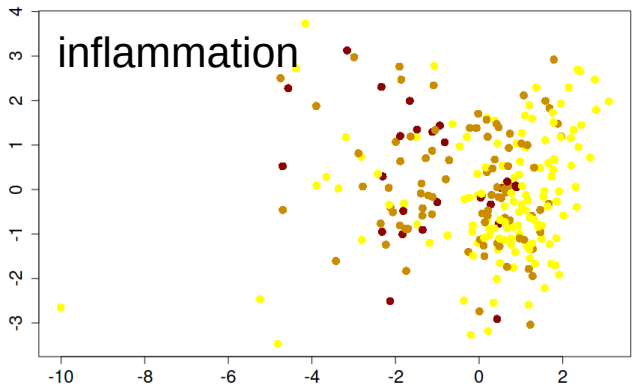
Clinical data PCA

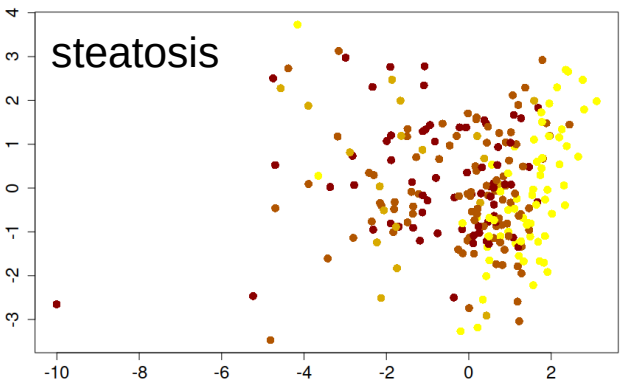




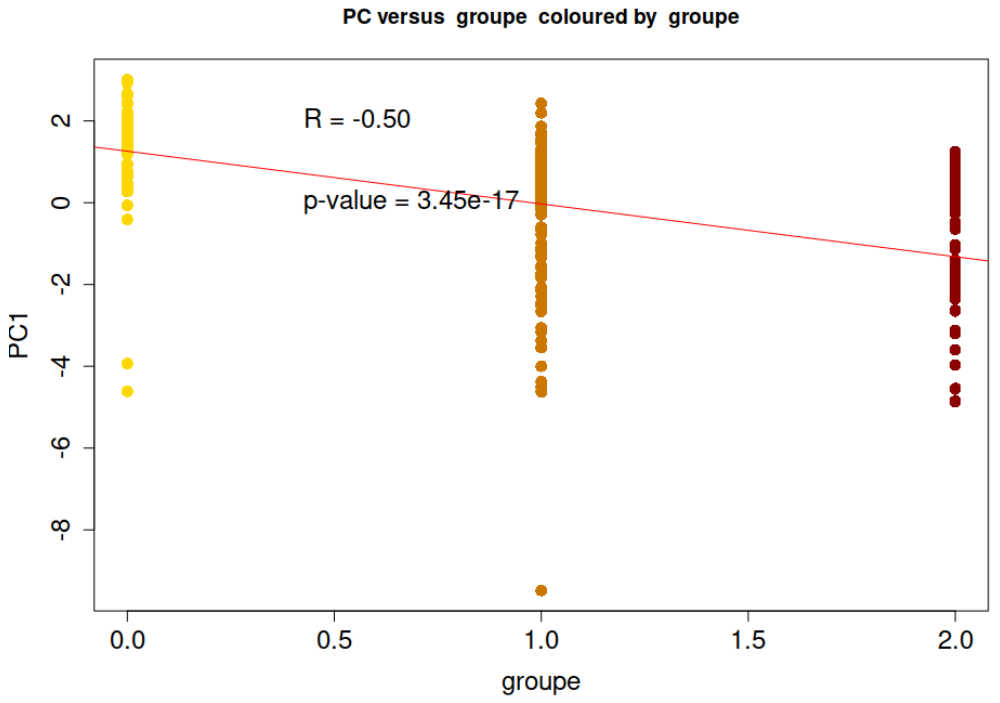
All clinical hallmarks align with PC1







What are the main loadings?

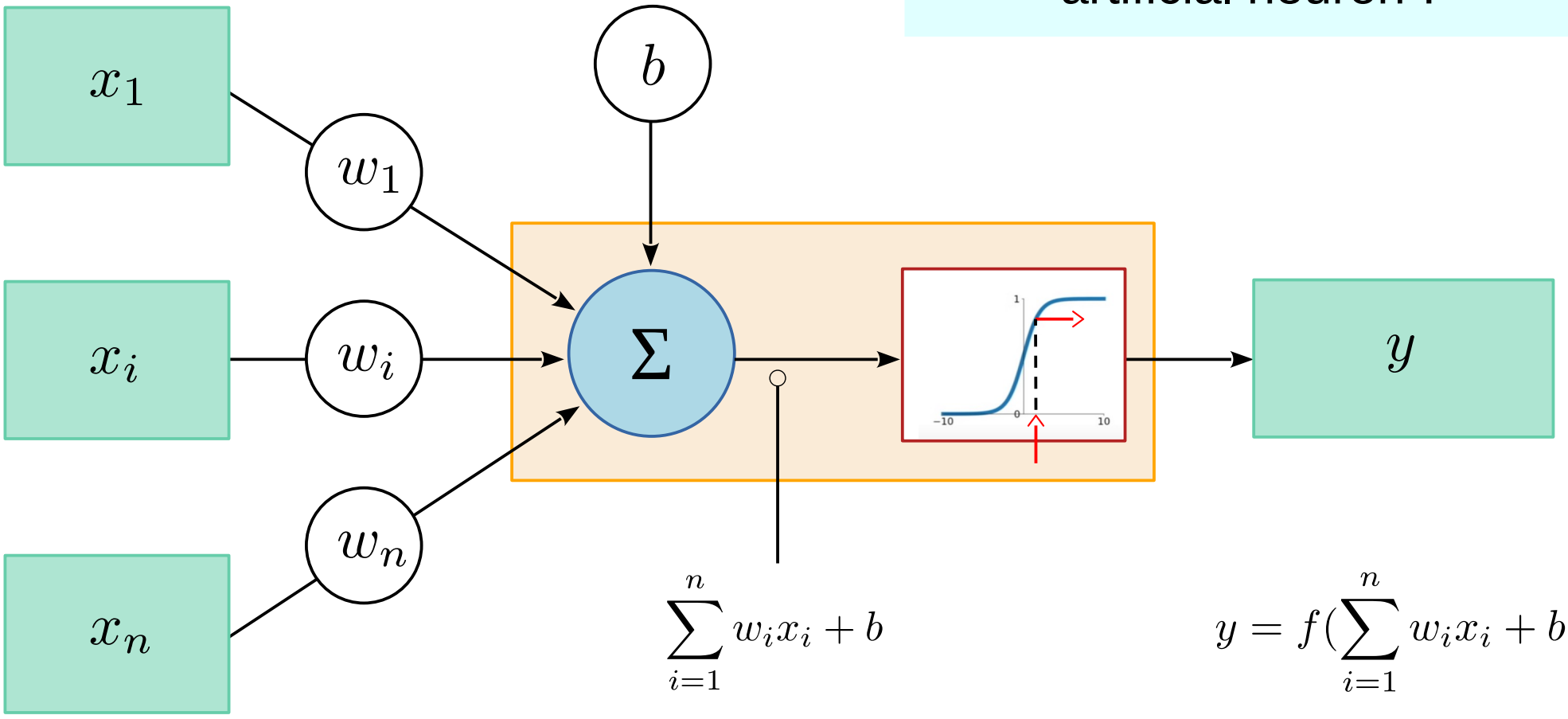


	PC1
M0_hba1c	-0.436
M0_glycemie_jeun_moy	-0.413
M0_gammaGT	-0.353
M0_triglycerides	-0.345
M0_ASAT_TGO	-0.306
M0_ALAT_TGP	-0.294
M0_age	-0.208
M0_insuline_jeun2	-0.168
M0_alpha2_macroglobuline	-0.136
M0_lymphocytes	-0.084
M0_bilirubine_totale	-0.053
M0_CrPus	0.007
M0_haptoglobine	0.032
sexe	0.074
M0_LDL	0.101
M0_apoA1	0.103
M0_plaquettes	0.195
M0_HDL	0.224

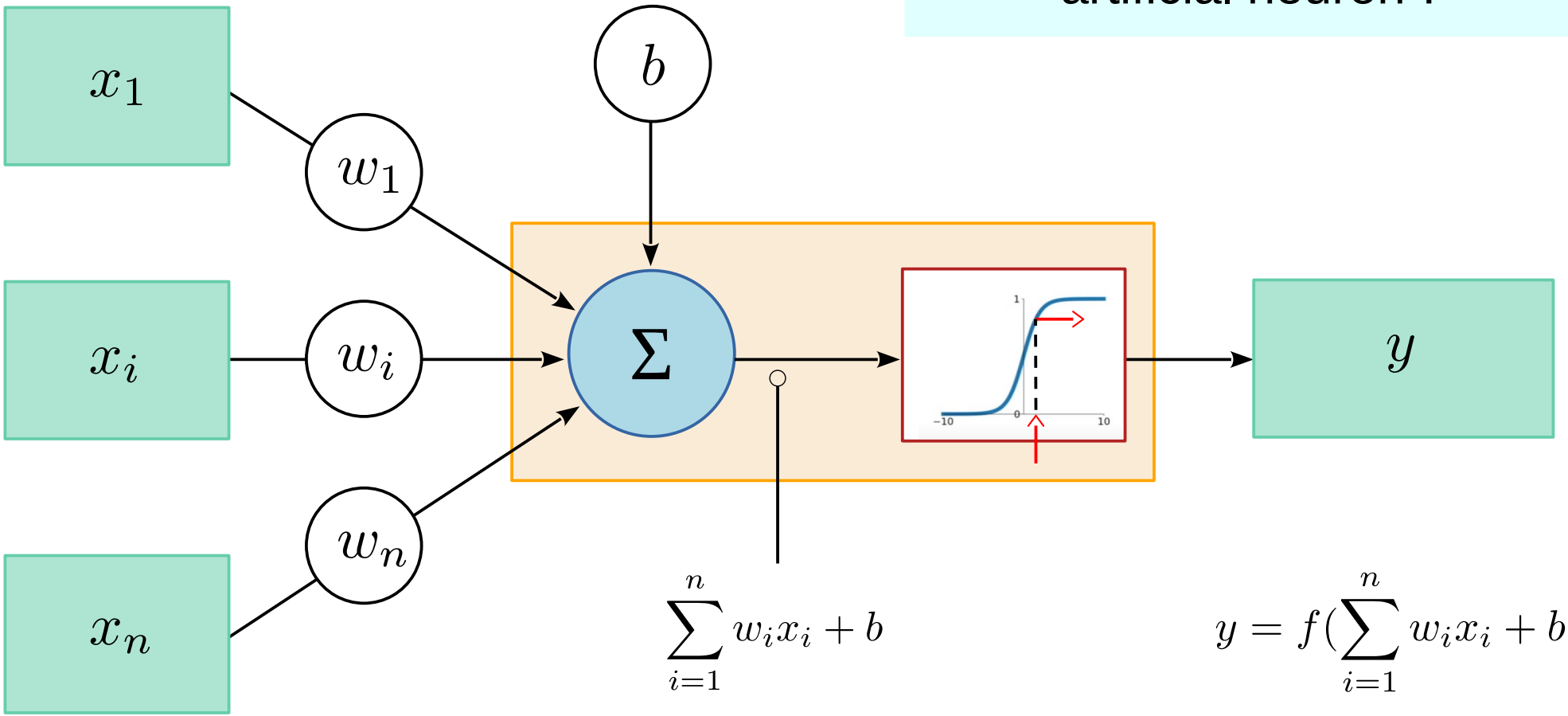
Can we train artificial neural networks

to recognise the presence of NAFL and NASH?

What is an artificial neuron ?



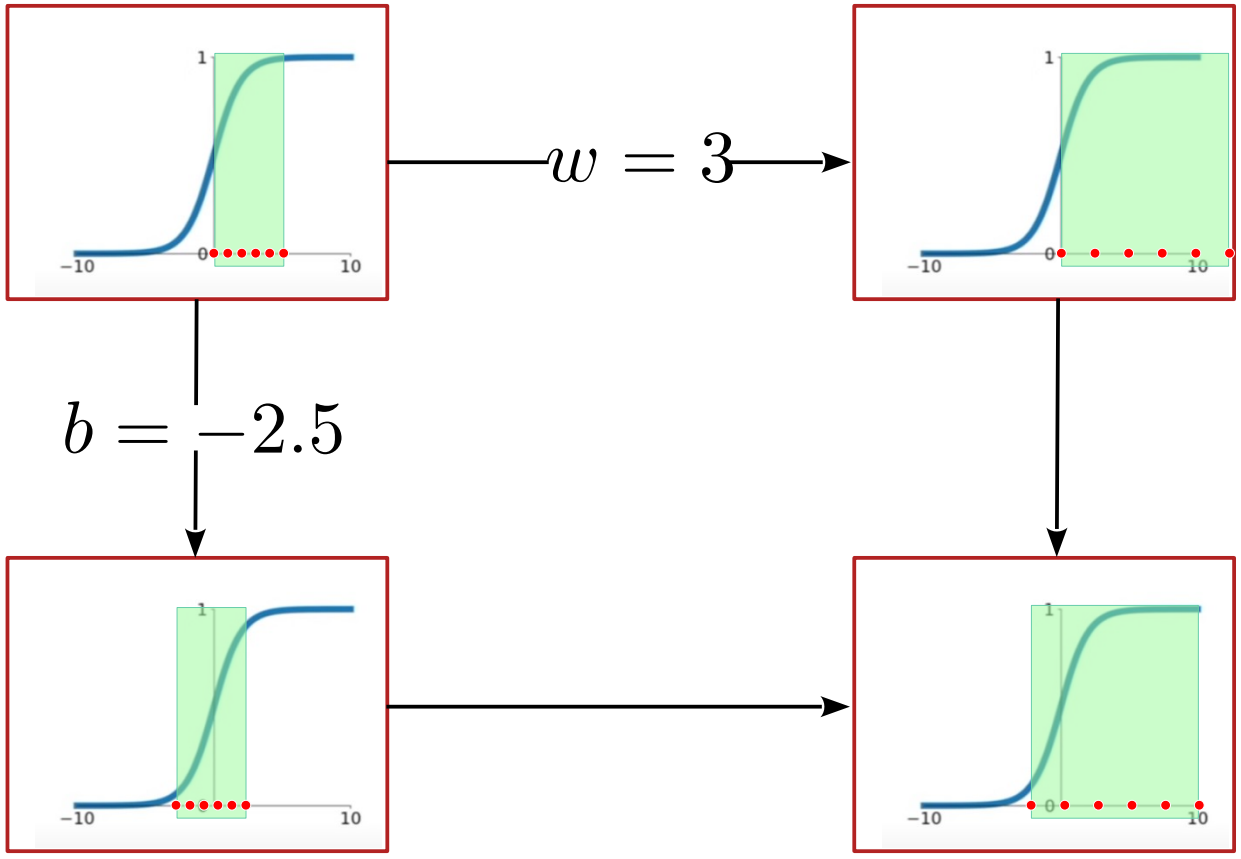
What is an artificial neuron ?



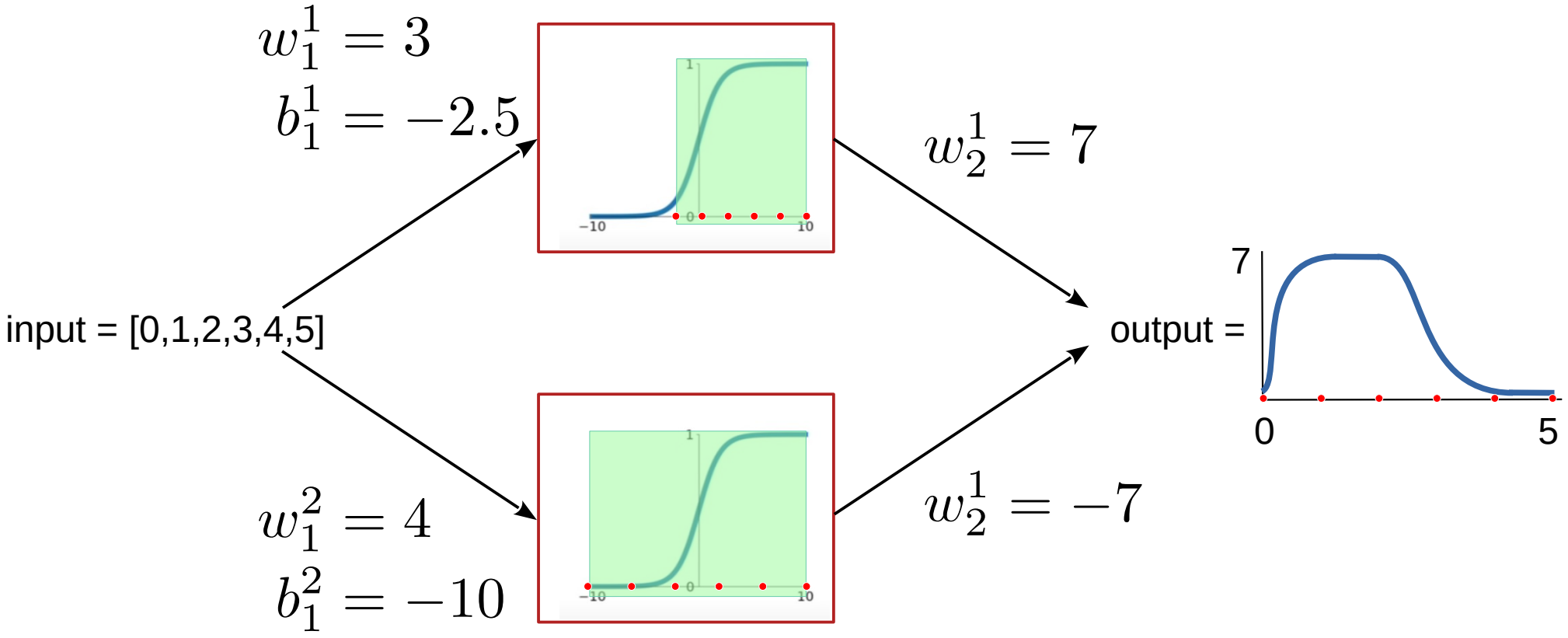
NB: when the activation function is logistic (sigmoid), this is actually a logistic regression...

Impact of the weights and the bias

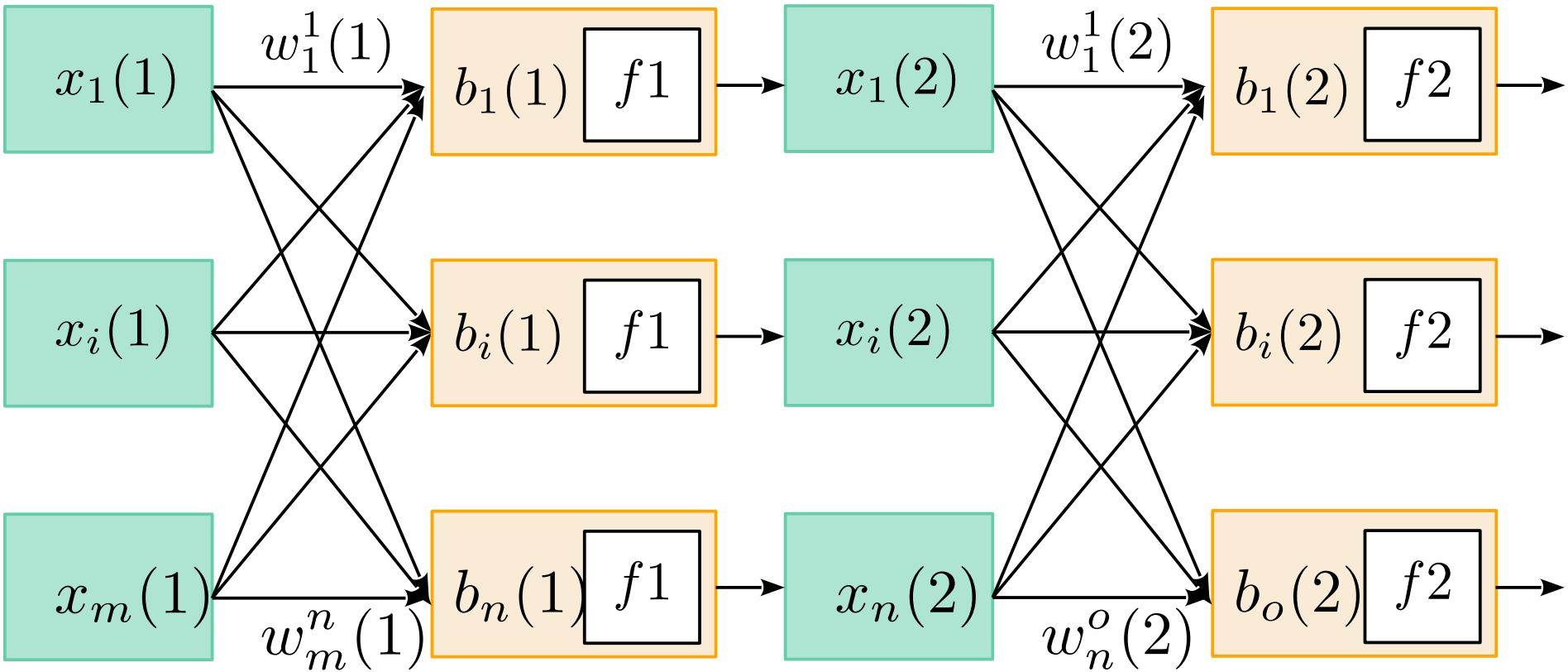
input = [0,1,2,3,4,5]



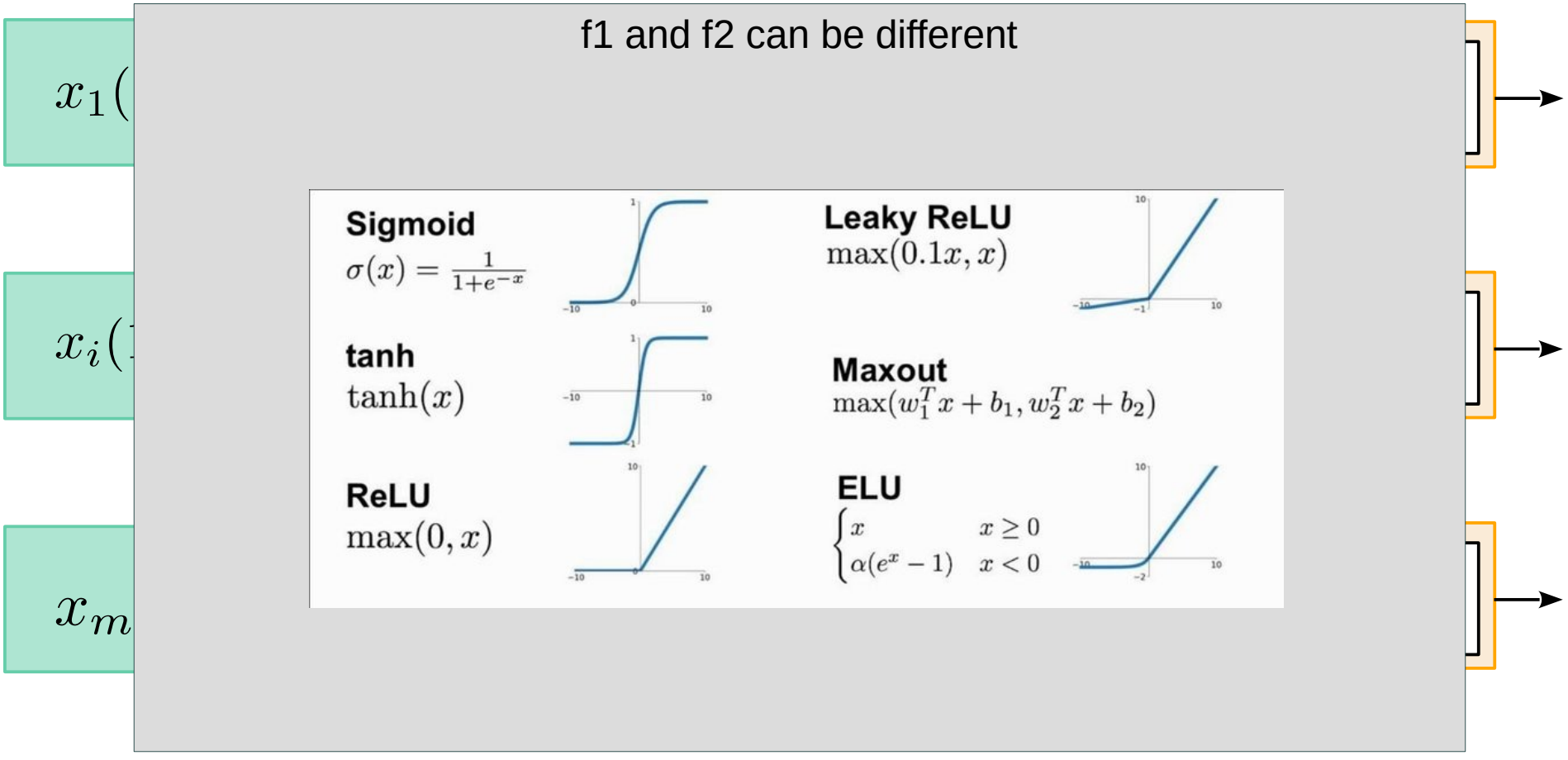
The magic happens with several neurons



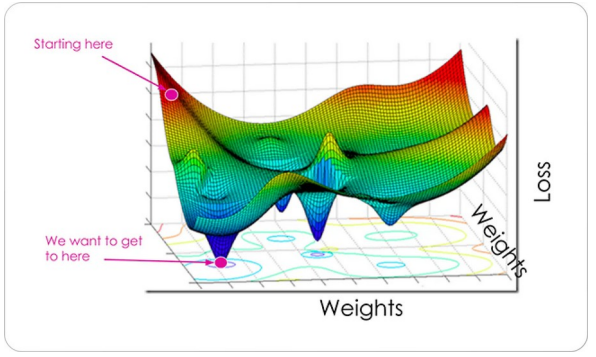
And then we add layers (the “Deep”)



Many different activation functions



Deep “learning”



Update of parameters

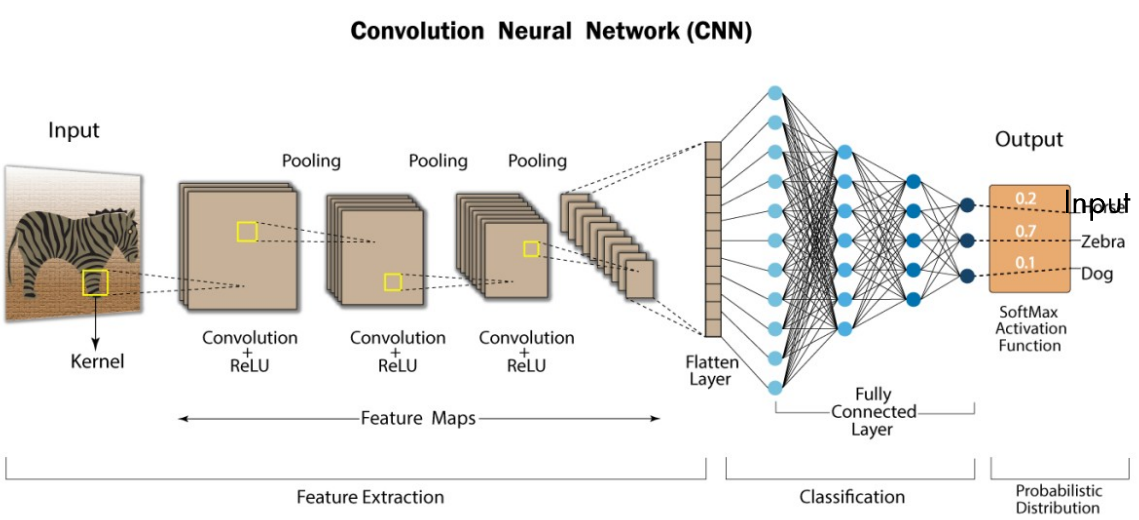
Minimising loss function

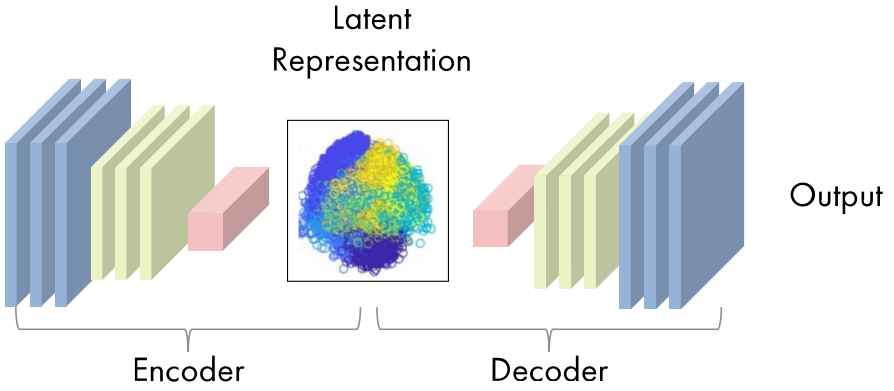
training set

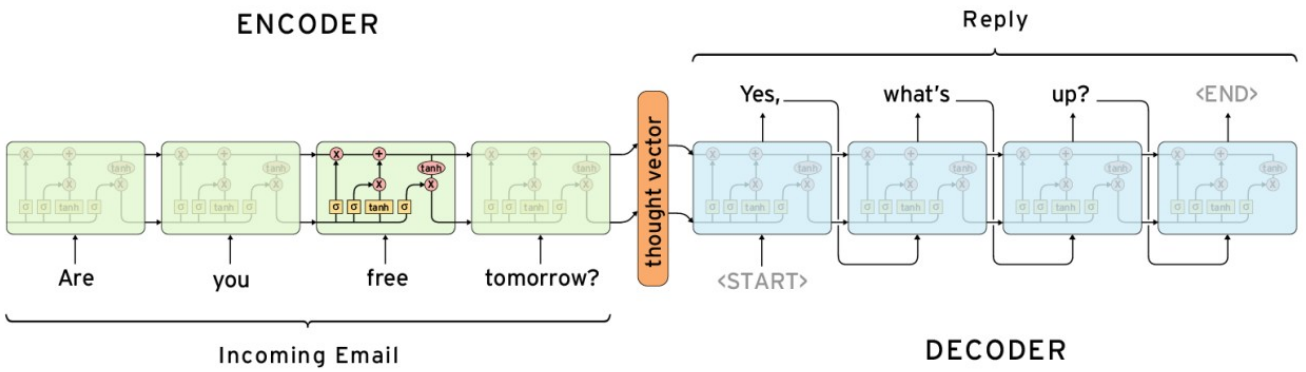
one set of inputs

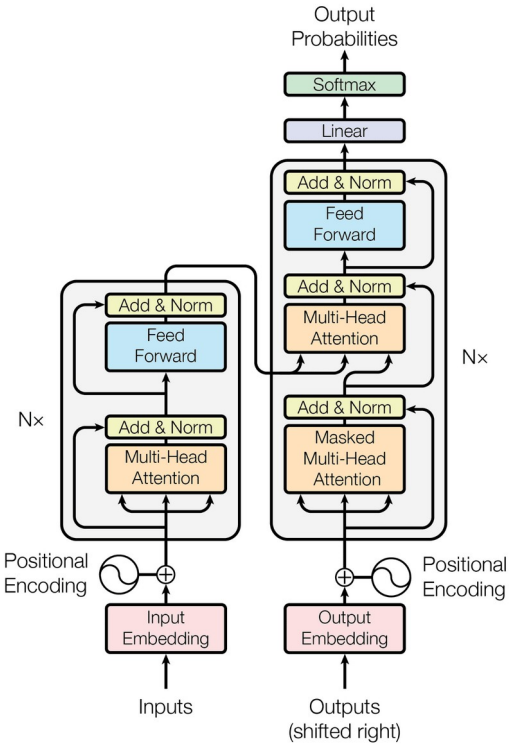
one set of true outputs

comparison


Convolution Neural Network (CNN)
Input
Pooling
Pooling
Pooling
Output
0.2
0.7
0.1
Horse
Zebra
Dog
Kernel
Convolution + ReLU
Convolution + ReLU
Convolution + ReLU
Flatten Layer
Fully Connected Layer
SoftMax Activation Function
Feature Maps
Feature Extraction
Classification
Probabilistic Distribution

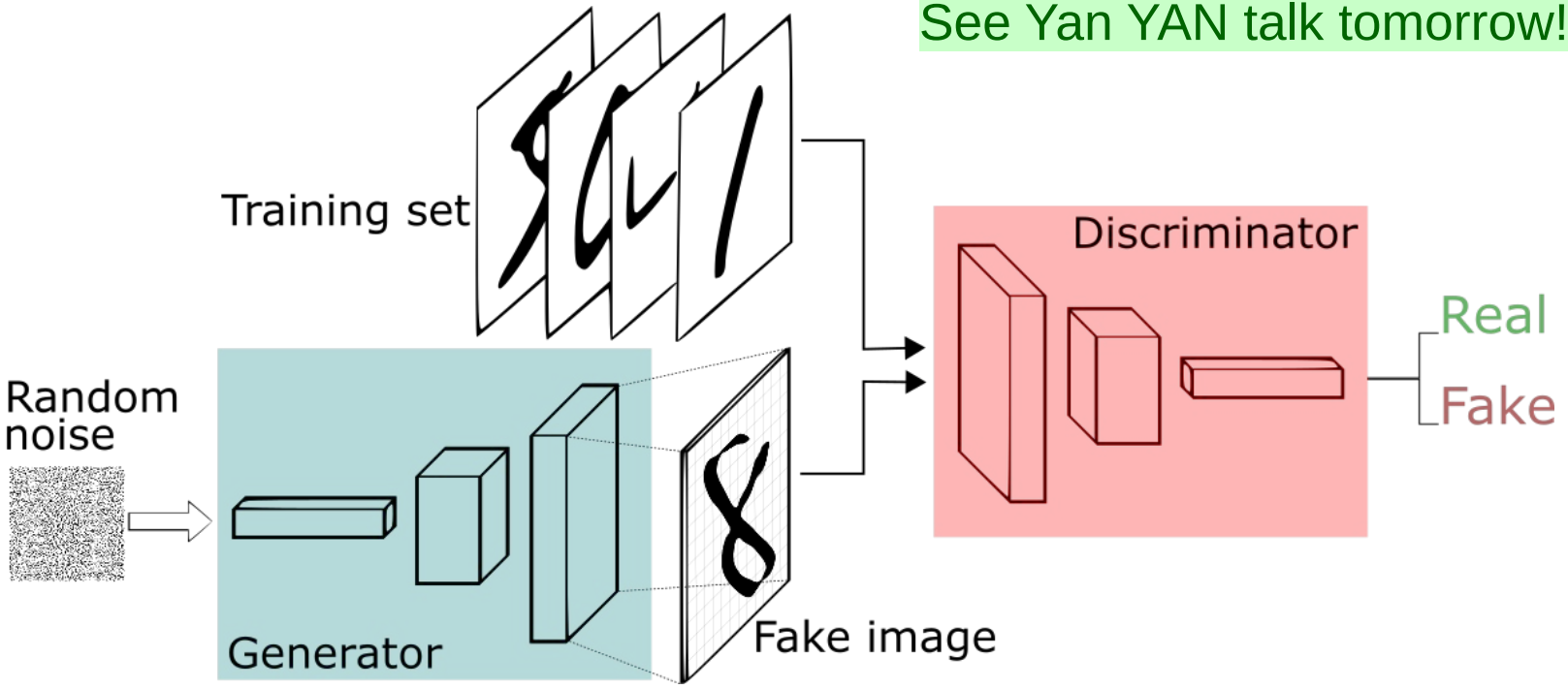

Latent Representation
Input
Output
Encoder
Decoder


ENCODER
Reply
Yes,
what's
up?
<END>
thought vector
Are
you
free
tomorrow?
<START>
Incoming Email
DECODER

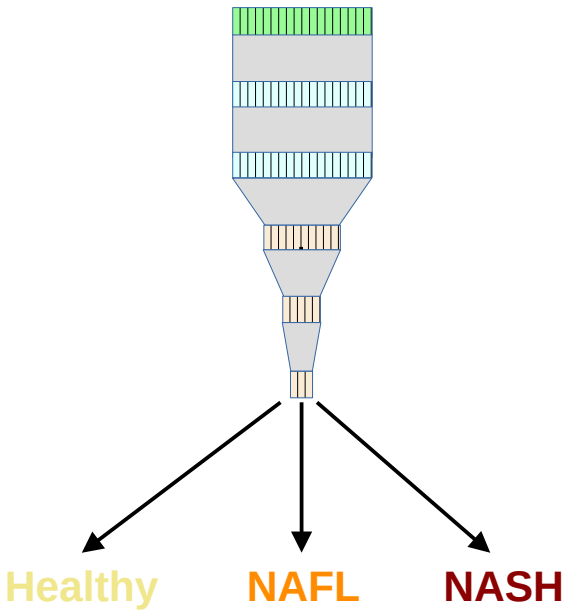

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Learning by trying to trick itself: Generative Adversarial Network (GAN)

See Yan YAN talk tomorrow!



Multi-Layer Perceptron trained on clinical data



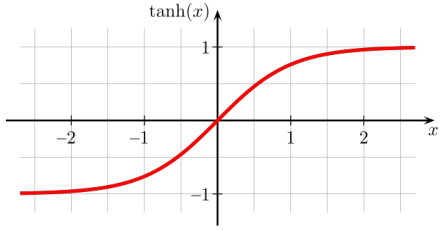
16 clinical data
+ age + sex

Dropout 30%

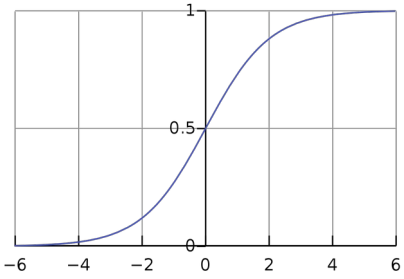
Randomly omitted inputs
at each learning step (epoch)

Normalisation

10 tanh



5 sigmoid

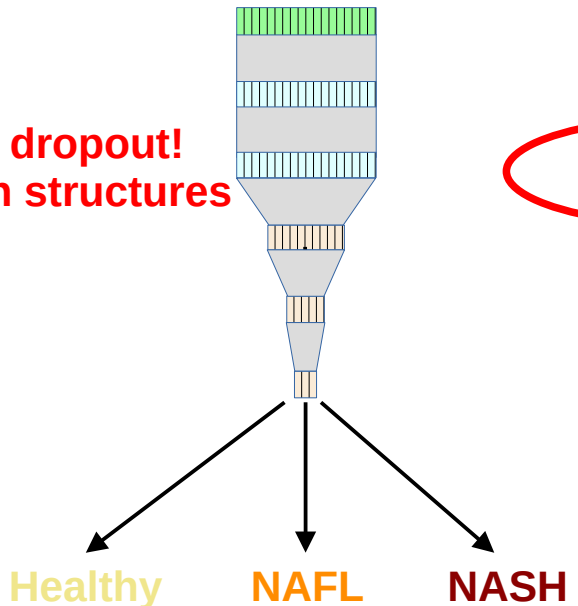


3 softmax

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$ = pseudo probability

Multi-Layer Perceptron trained on clinical data

After input and dropout!
→ avoid hidden structures



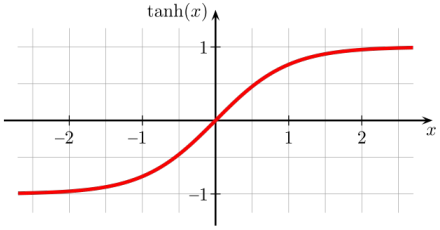
16 clinical data
+ age + sex

Dropout 30%

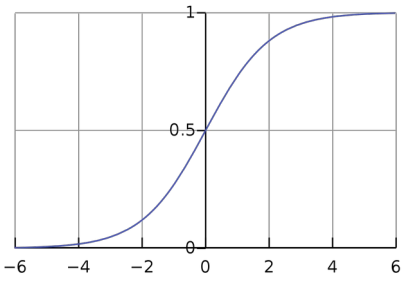
Randomly omitted inputs
at each learning step (epoch)

Normalisation

10 tanh



5 sigmoid



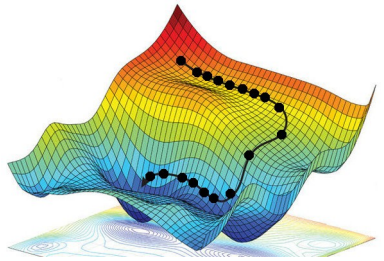
3 softmax

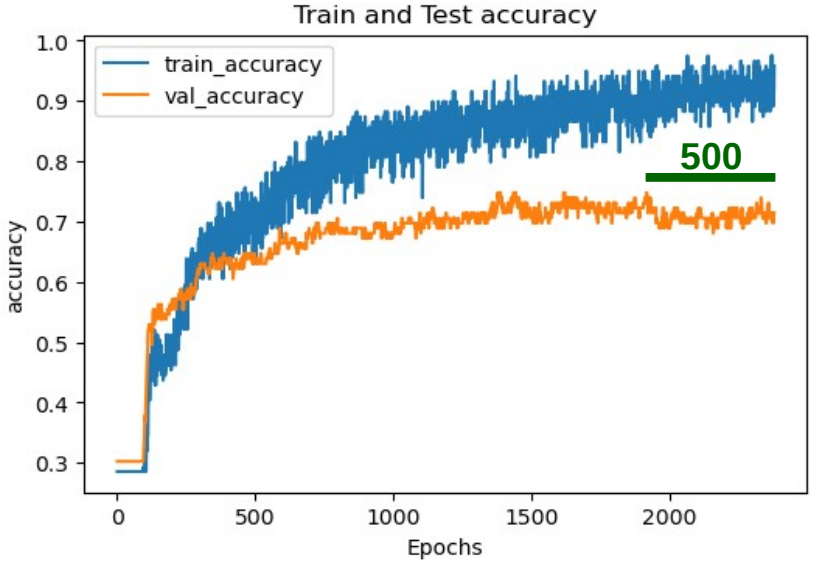
$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$ = pseudo probability

Hyperparameters (fancy word for settings)

- Evaluation
  - Loss function: categorical cross-entropy
  - Optimizer: Adam (learning rate = 0.001)
- Training duration
  - Epochs = 10000
  - Batch size = 16
- Early stop:
  - min delta = 0.001,
  - patience = 500 (1000)

$$-\sum_i truth_i \times log(pred_i)$$





Training procedure

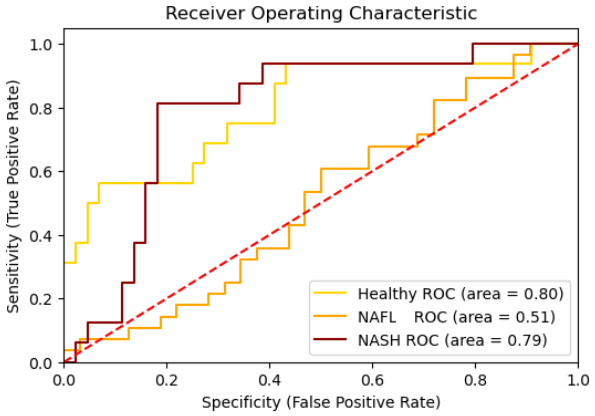
5 independently trained models

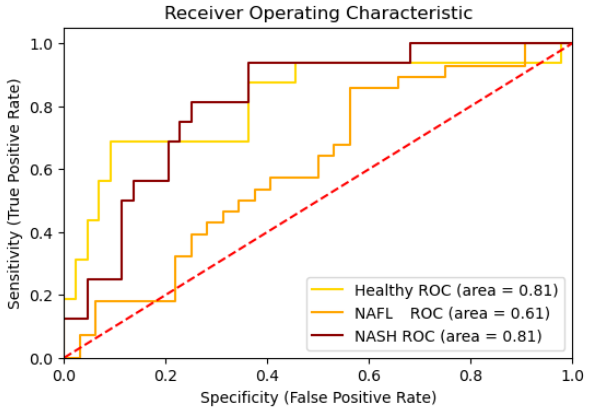
Validation (never seen): 60 subjects
Same for all model instances

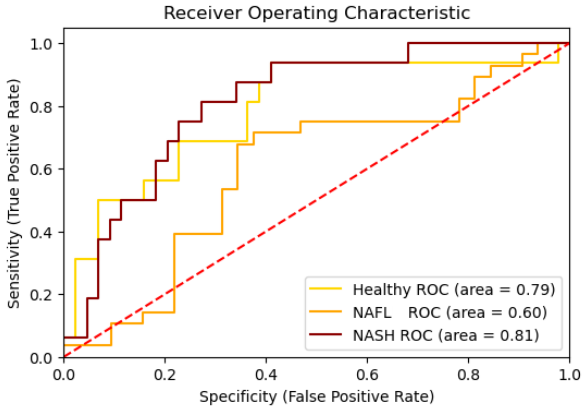
Training set: 178 subjects

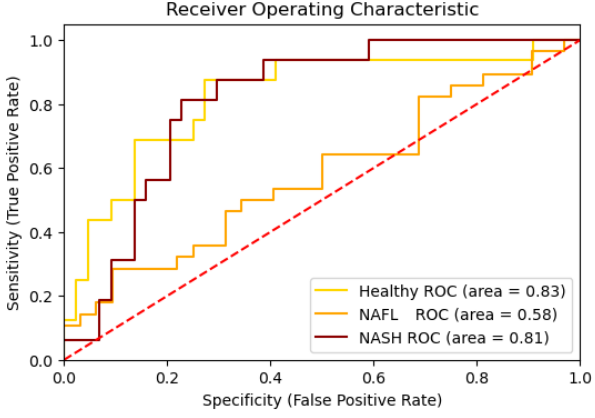
Test set (for training): 60 subjects
Different for each model instance

Clinical data predictivity on the independent dataset

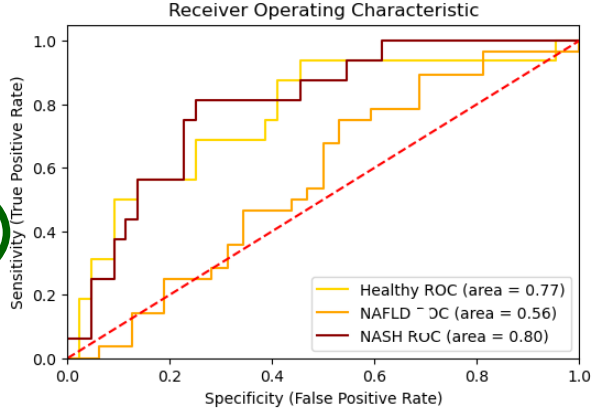




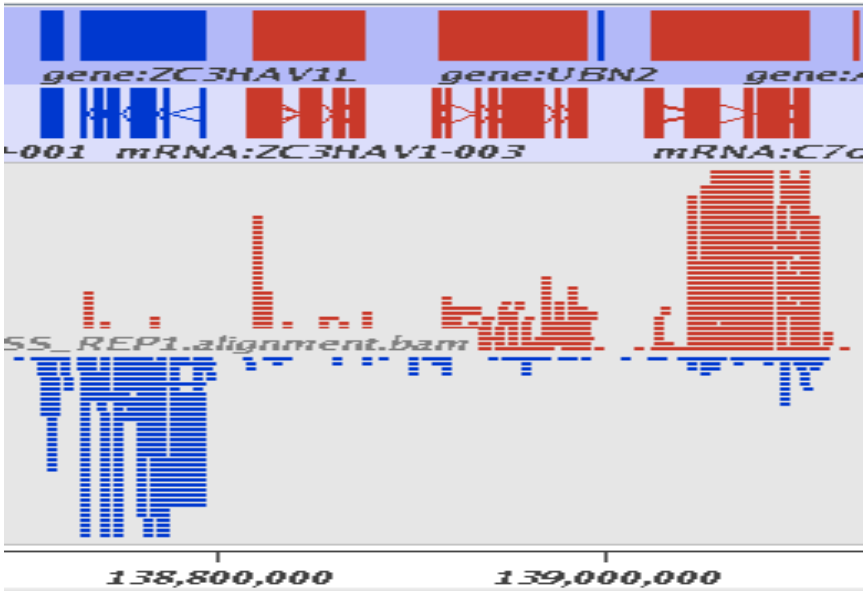




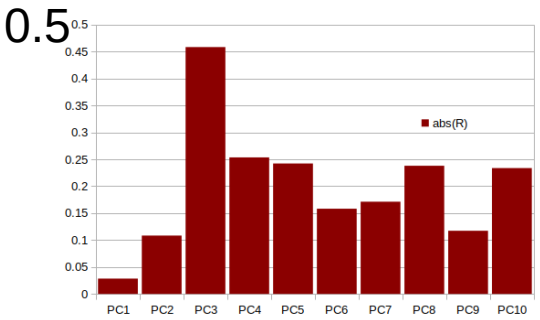
	healthy	NAFL	NASH	overall
TP	11	17	8	
TN	39	20	37	
FP	5	12	7	
FN	5	11	8	
AUC	0.79	0.62	0.67	0.68
accuracy	83.33	61.67	75.00	60.00
sensitivity	88.64	62.50	84.09	78.41
specificity	68.75	60.71	50.00	59.82
precision	68.75	58.62	53.33	60.23

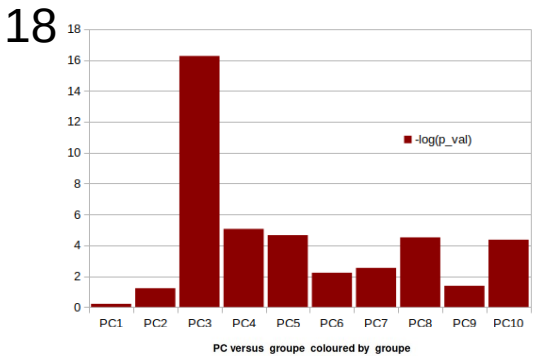


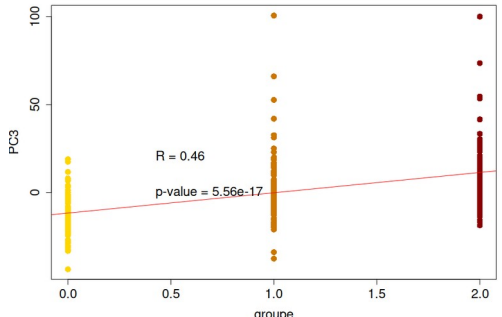
Transcriptomics

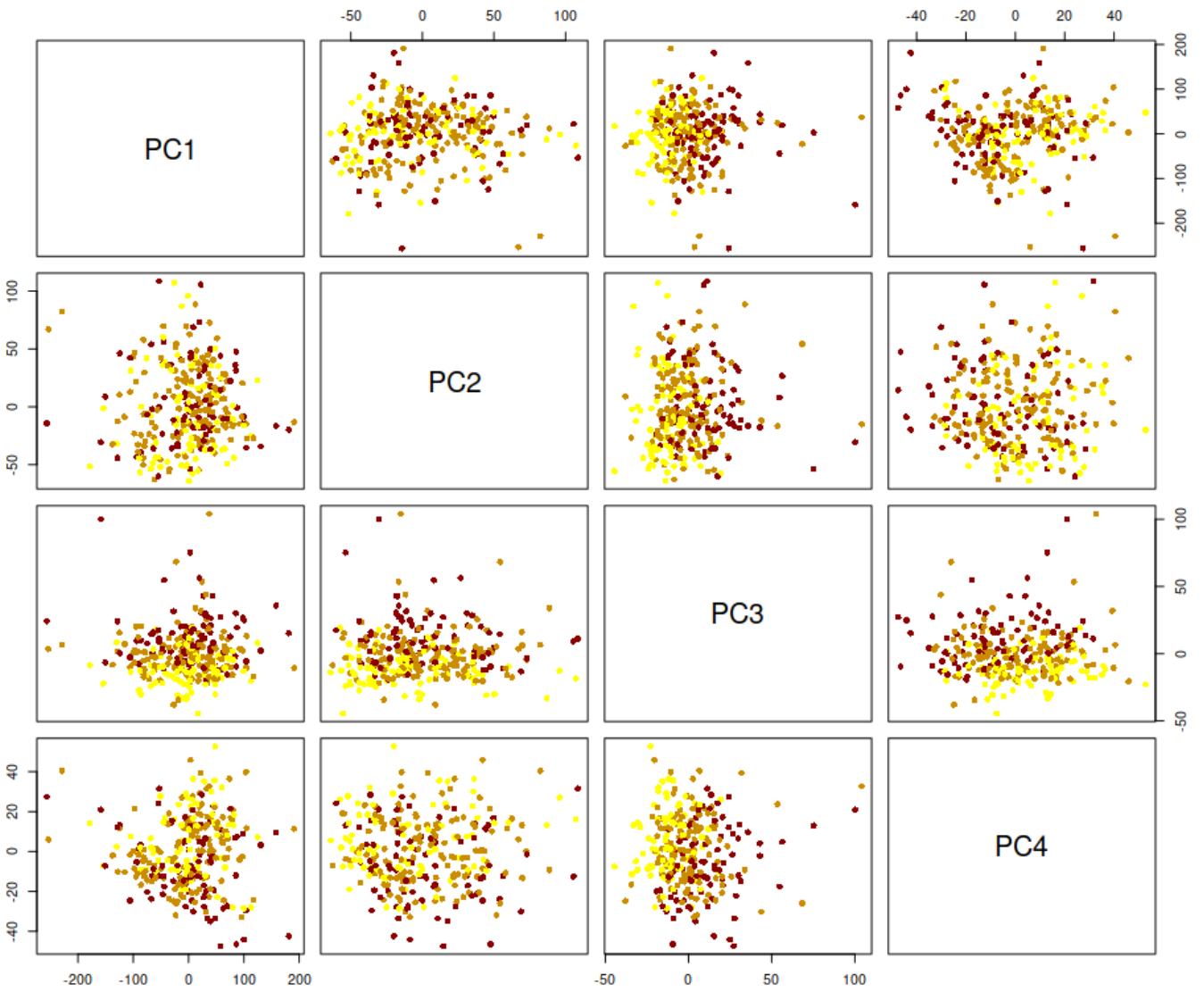


PCA RNAseq coloured by group



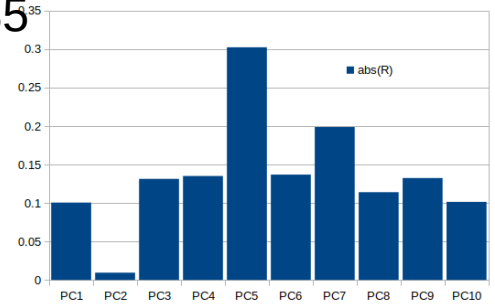




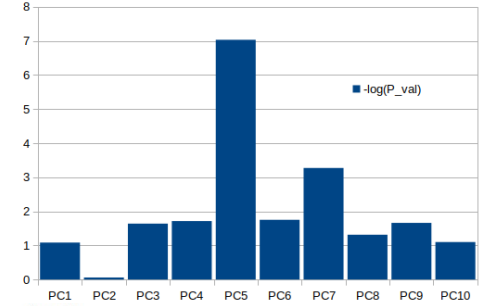


PCA RNAseq coloured by age

0.35



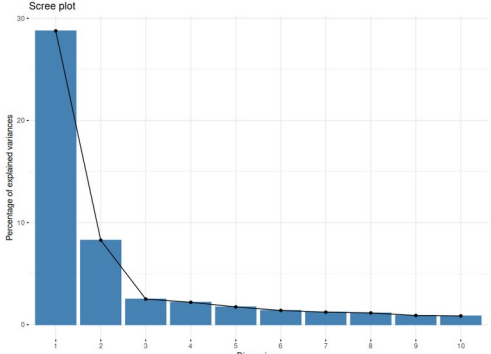
8

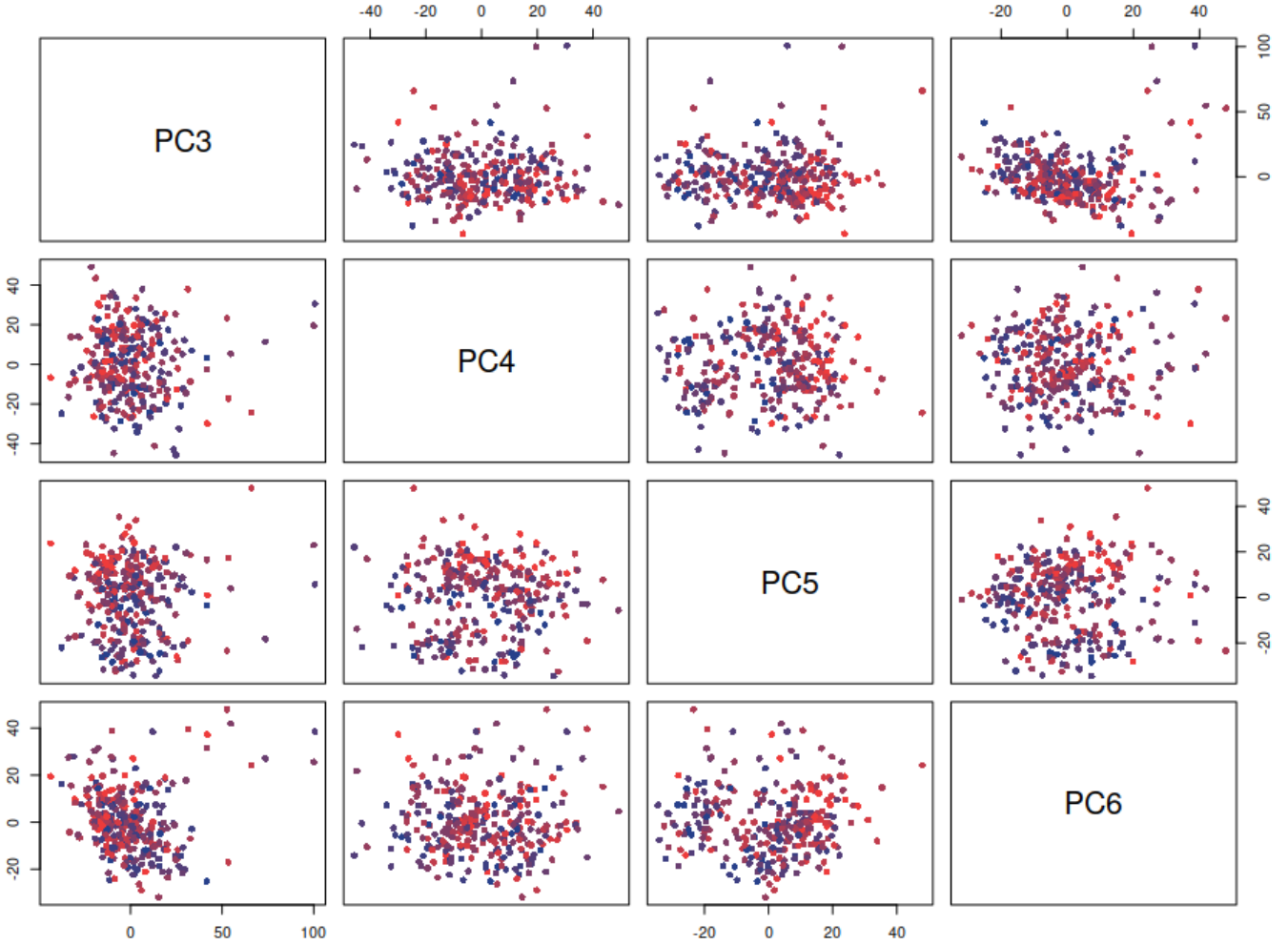


30%

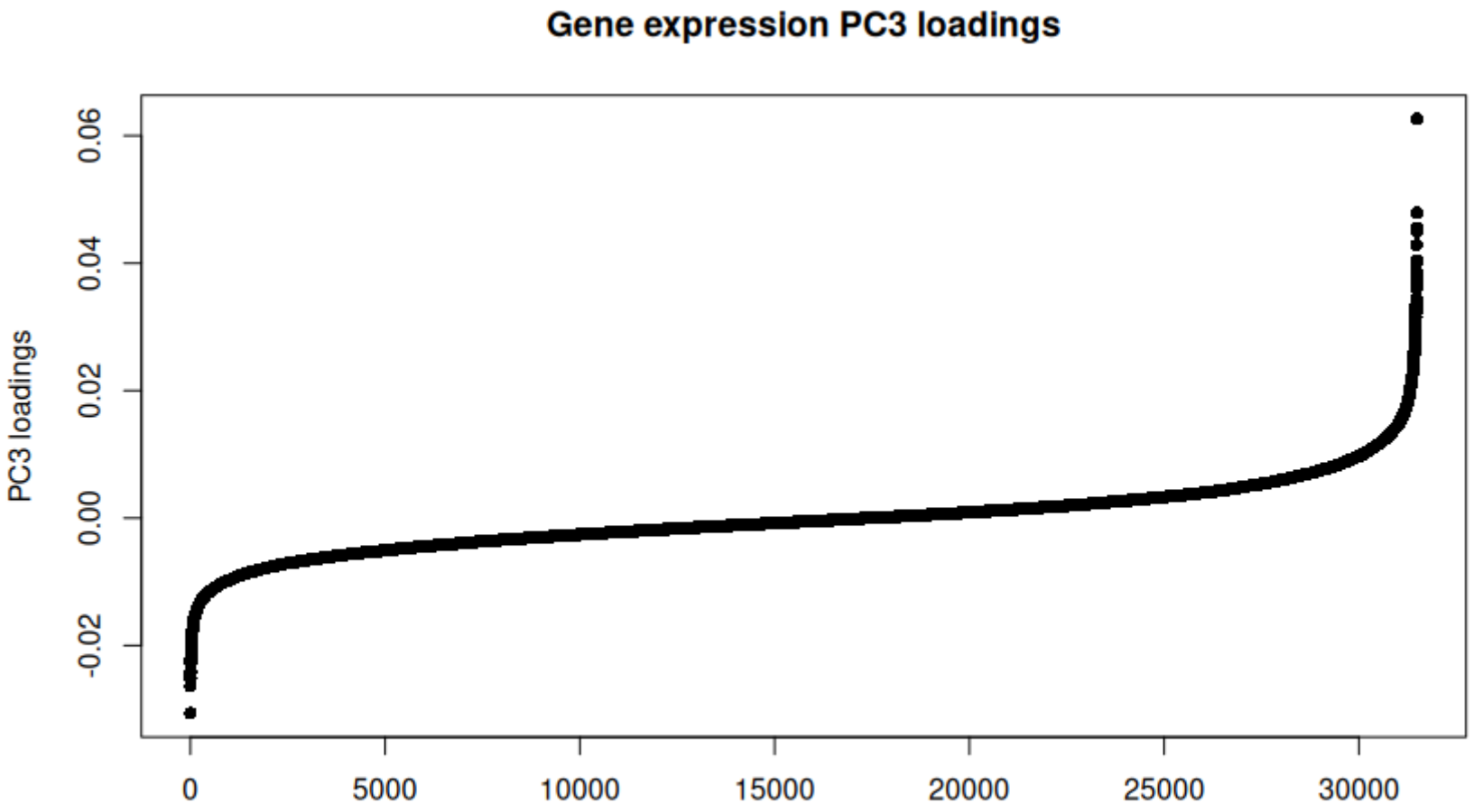
20%

10%

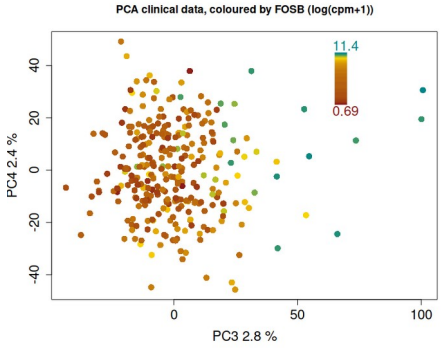


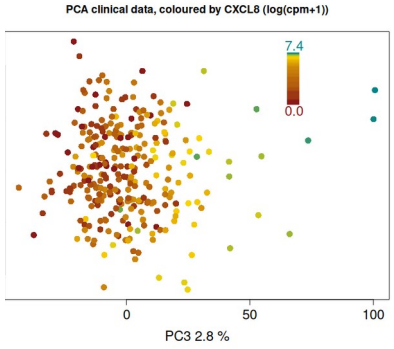


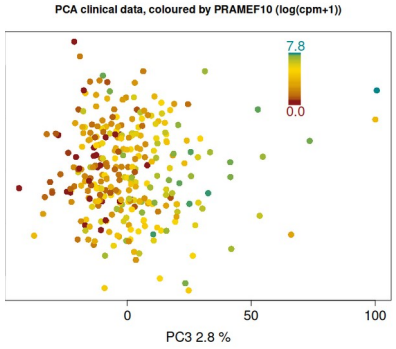
Alignment along PC3 controlled by few genes

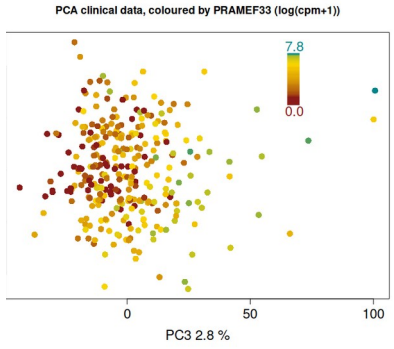


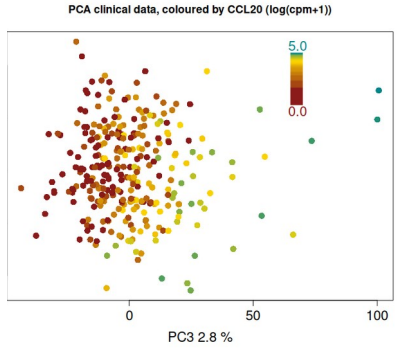
Main loadings (coloured by log[expression])

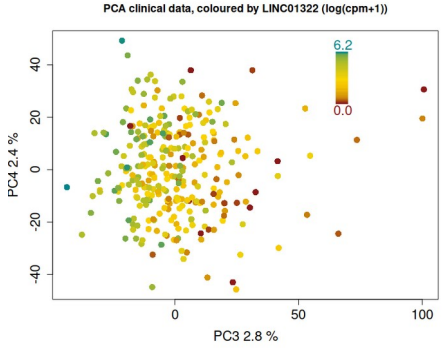


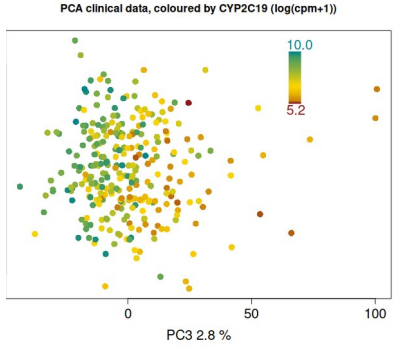


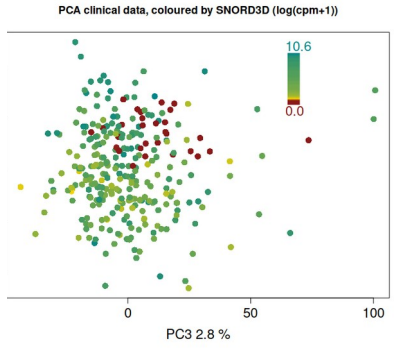


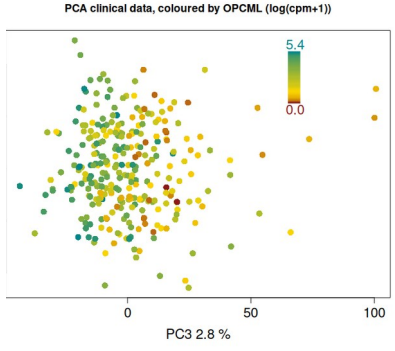




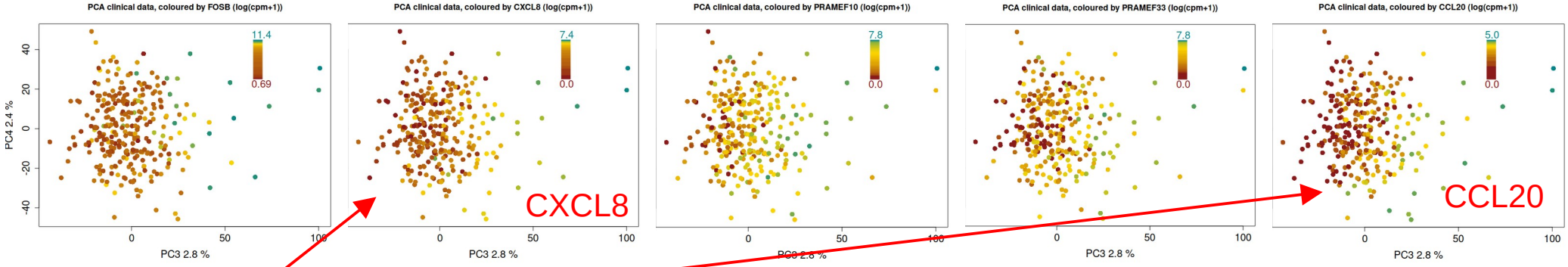




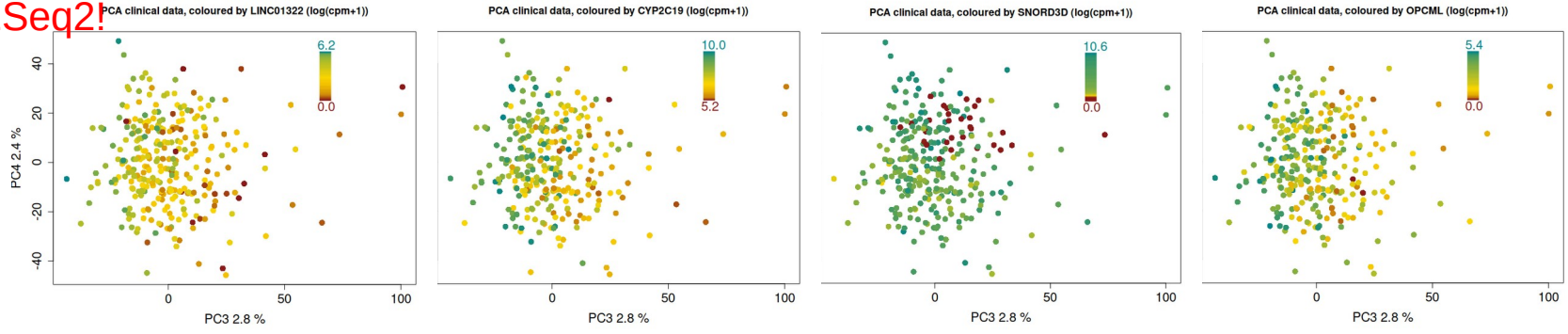




Main loadings (coloured by log[expression])



Not significant with DESeq2!



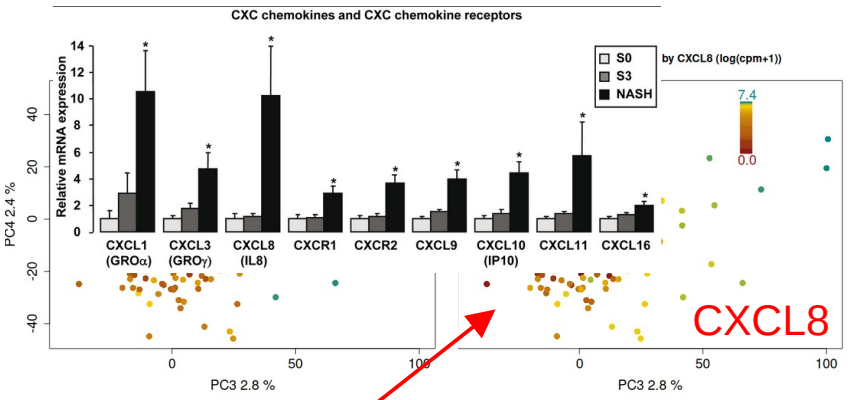
Hepatic Expression Patterns of Inflammatory and Immune Response Genes Associated with Obesity and NASH in Morbidly Obese Patients

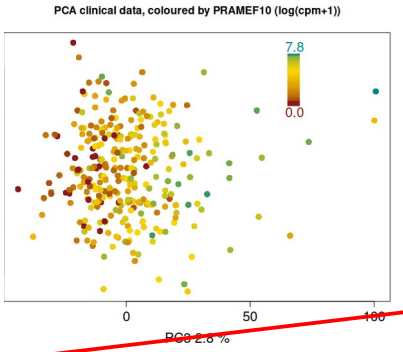
Adeline Bertola, Stéphanie Bonnafous, Rodolphe Anty, Stéphanie Patouraux, Marie-Christine Saint-Paul, Antonio Iannelli, Jean Gugenheim, Jonathan Barr, José M. Mato, Yannick Le Marchand-Brustel, Albert Tran, Philippe Gual

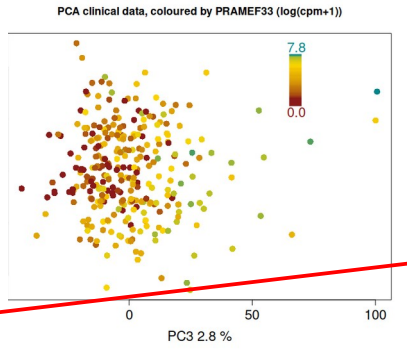
CCL20 IS A SENSITIVE AND SPECIFIC BIOMARKER FOR FIBROSIS RELATED TO NON-ALCOHOLIC STEATOHEPATITIS IN THE MORBIDLY OBESE

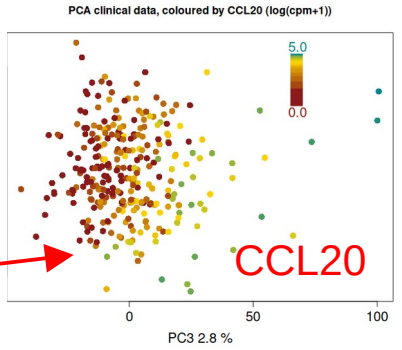
1016

Gerhard, Glenn S.[1,2]; Still, Christopher D.[1]; Petrick, Anthony T.[1]; Chu, Xin[1]; Shi, Weixing[1]; Chen, Hui[2]

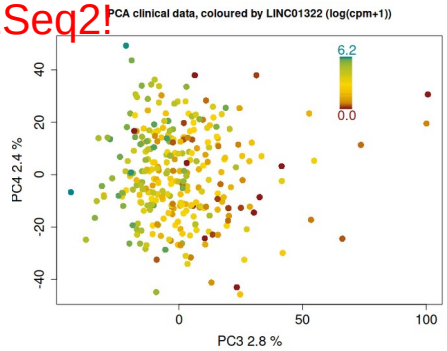


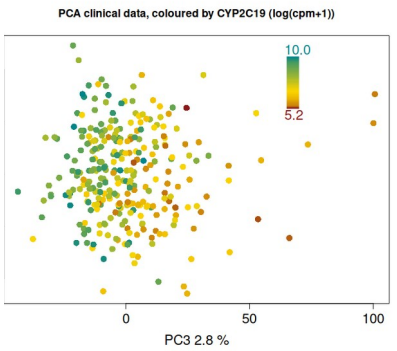


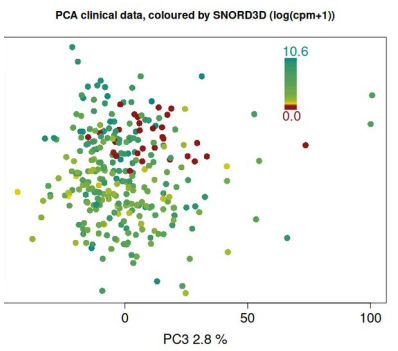


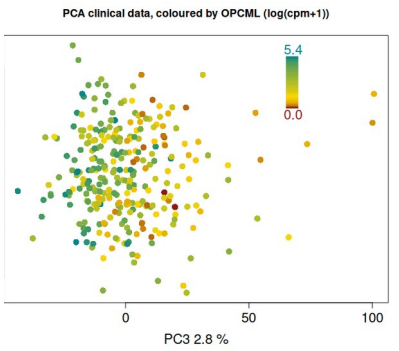


Not significant with DESeq2!



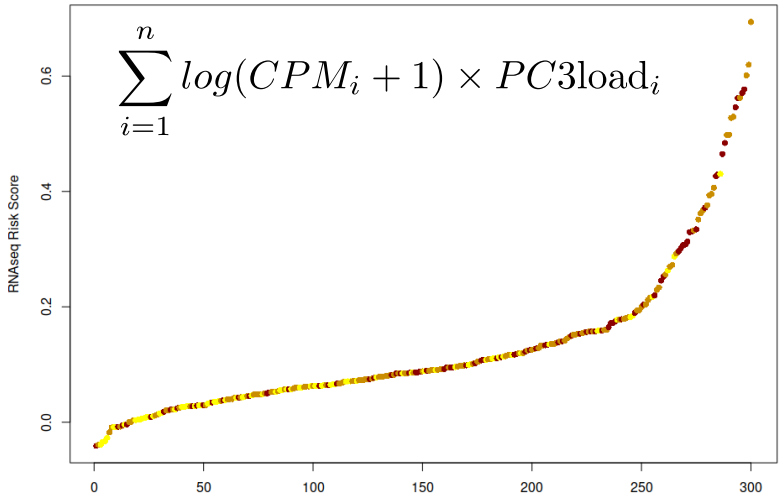




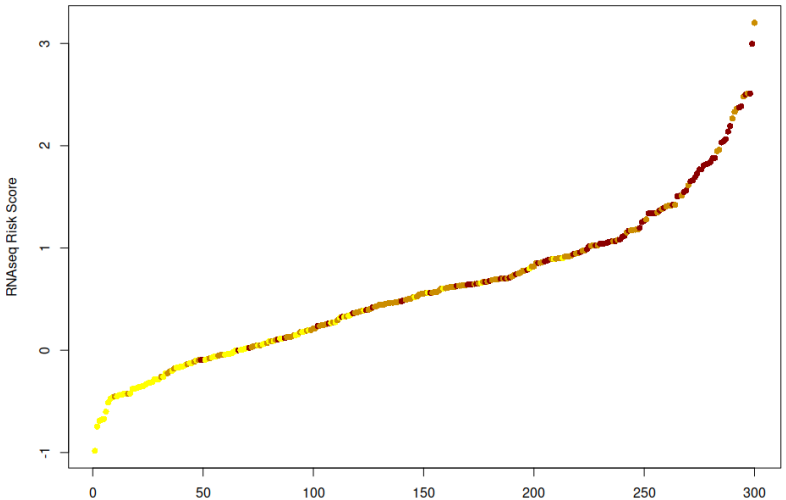


$i = 1 + 1$

$$\sum_{i=1}^{n} log(CPM_i + 1) \times PC3\text{load}_i$$

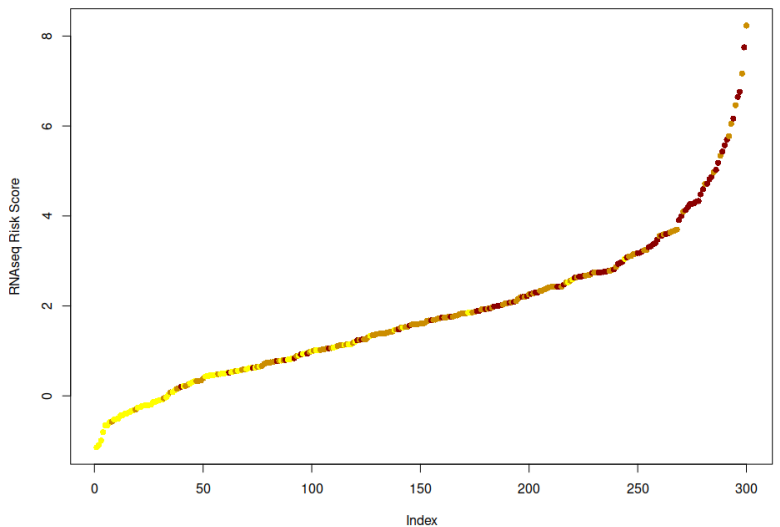


$i = 30 + 30$

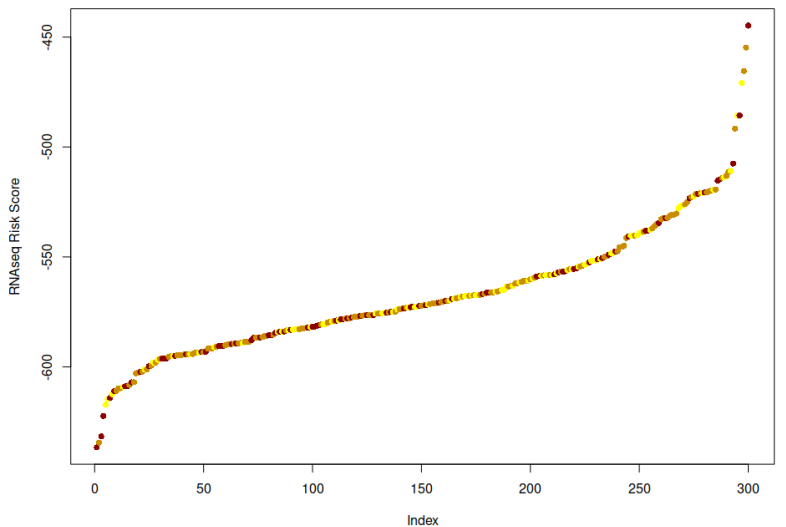


RNAseq Risk Score

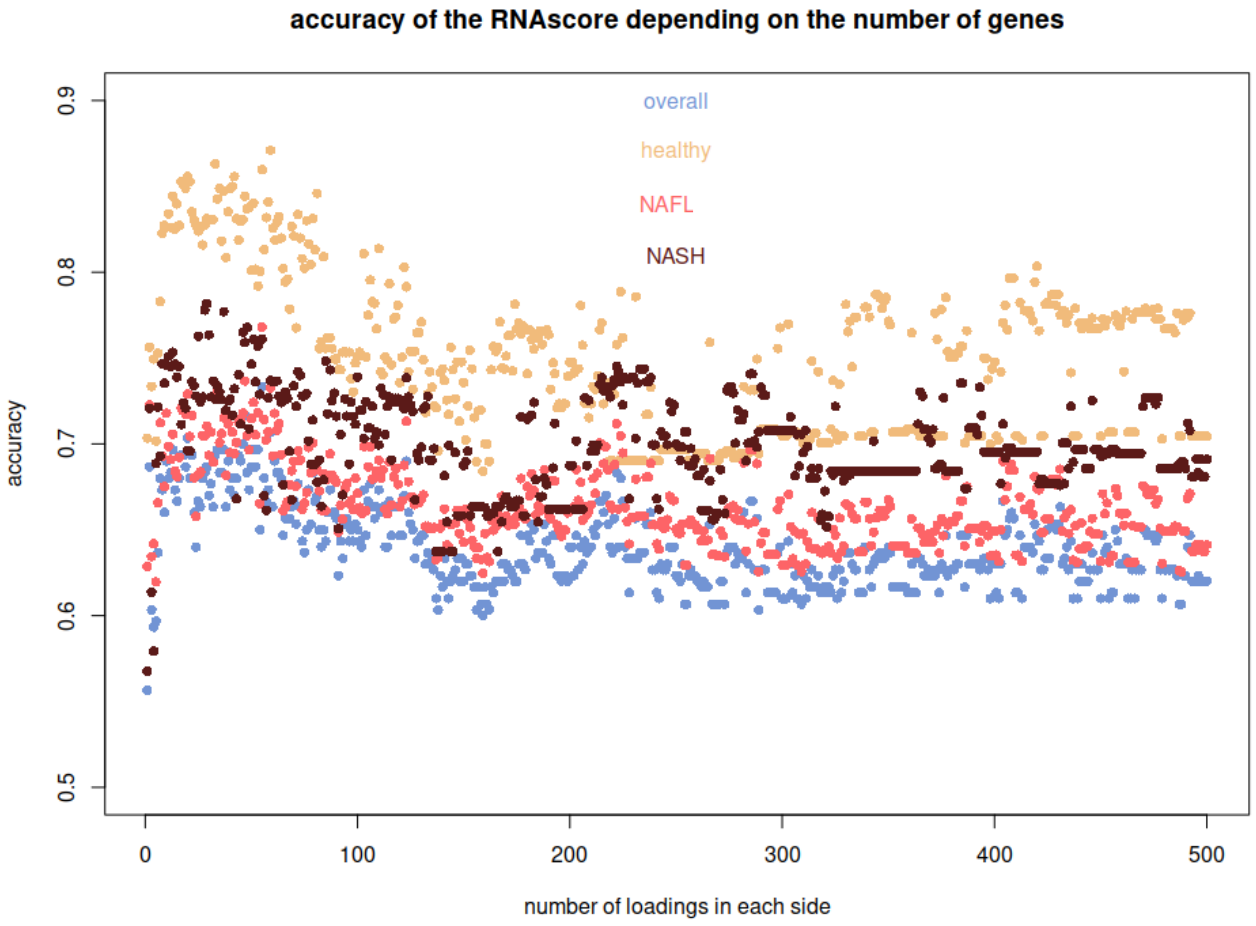
$i = 100 + 100$



All genes



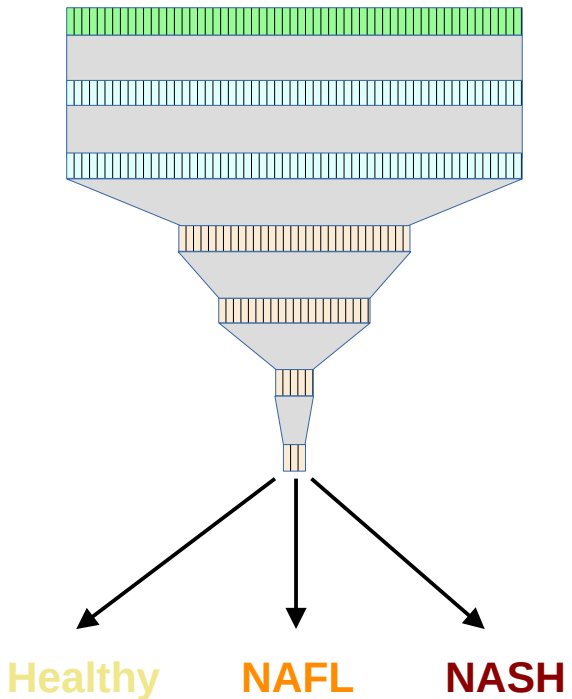
Are PC3 loadings predictive?



Random Forest
to choose cut-offs
between classes

Accuracy of the
classification made
on varying number
of loadings from each side

Multi-Layer Perceptron trained on RNAseq data



100+100 gene expression values (counts)

Dropout 30%

Normalisation

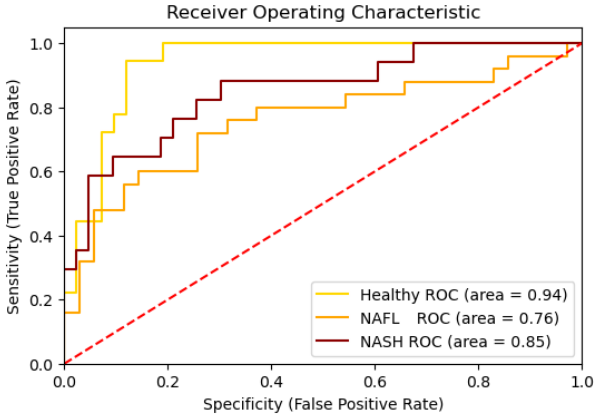
100 tanh

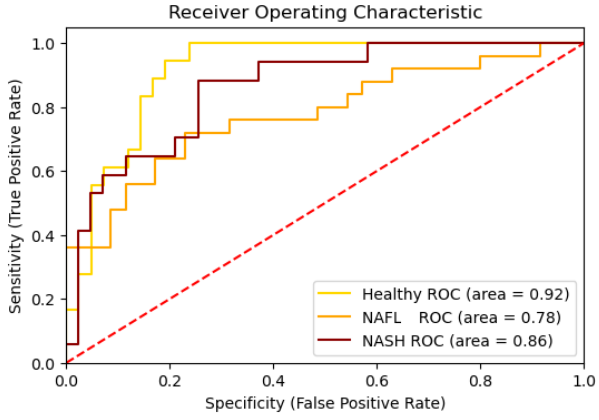
20 tanh

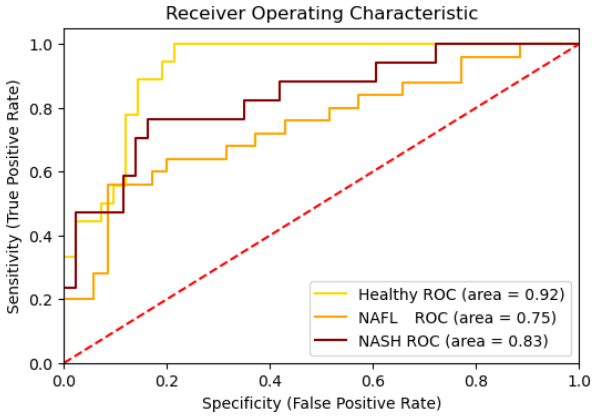
5 sigmoid

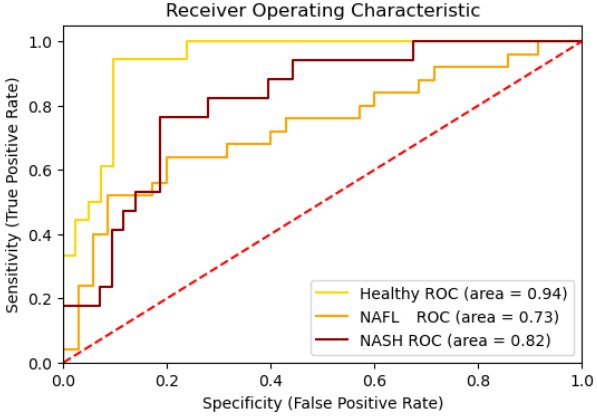
3 softmax

RNAseq data predictivity on the independent dataset

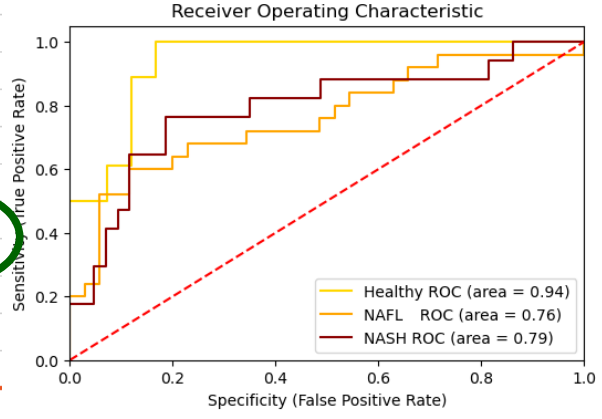




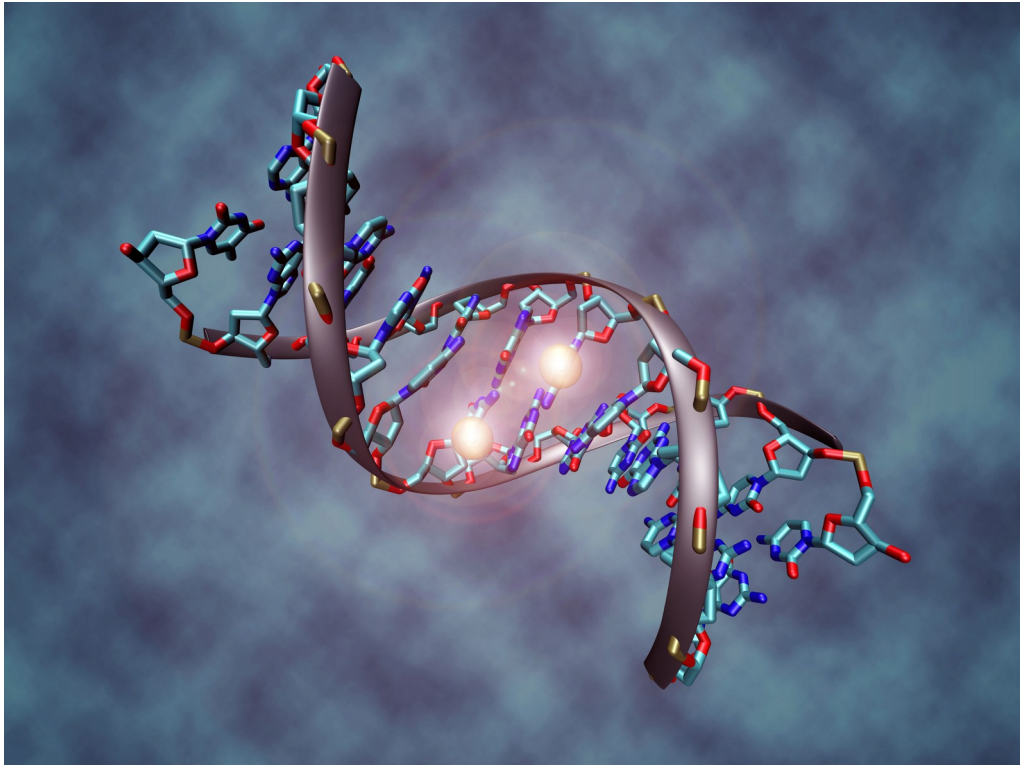




	healthy	NAFL	NASH	overall
TP	14	17	8	
TN	37	24	38	
FP	5	11	5	
FN	4	8	9	
AUC	0.83	0.68	0.68	0.73
accuracy	85.00	68.33	76.67	65.00
sensitivity	88.10	68.57	80.37	81.68
specificity	77.78	68.00	47.06	64.28
precision	73.68	60.71	61.54	65.31

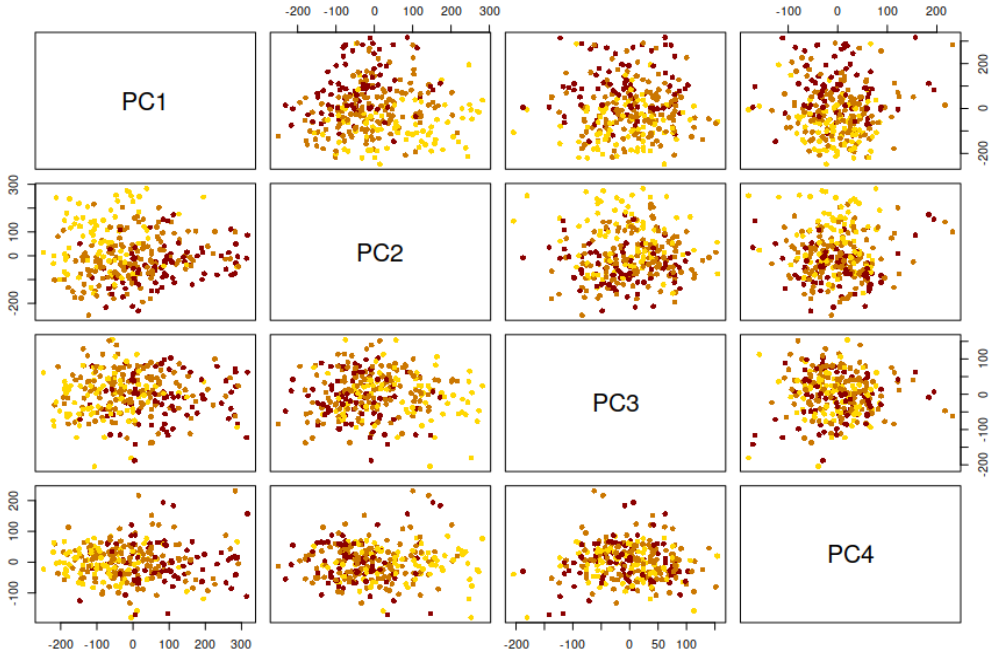


DNA Methylation

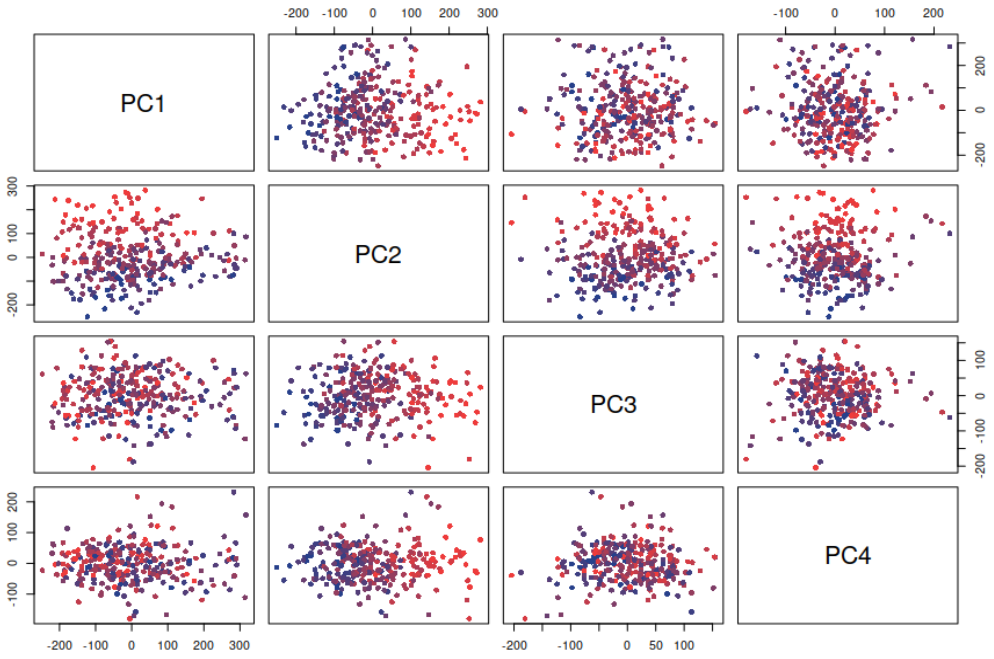


Severity and age both contribute to the main PCs

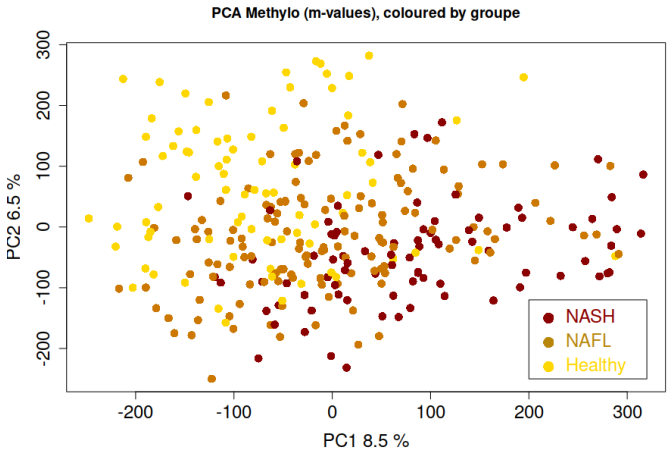
Coloured by severity

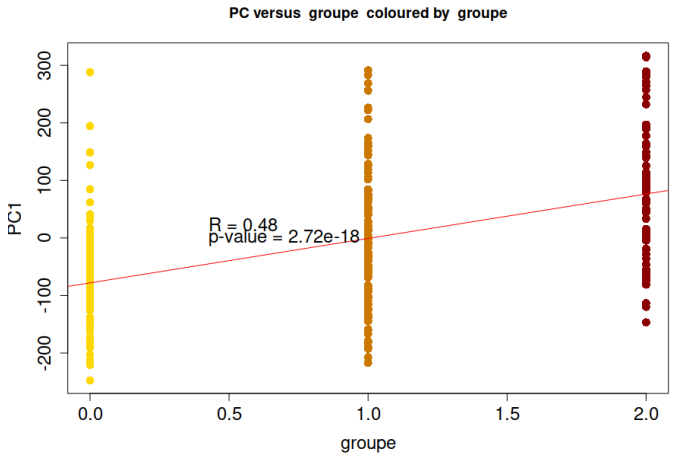


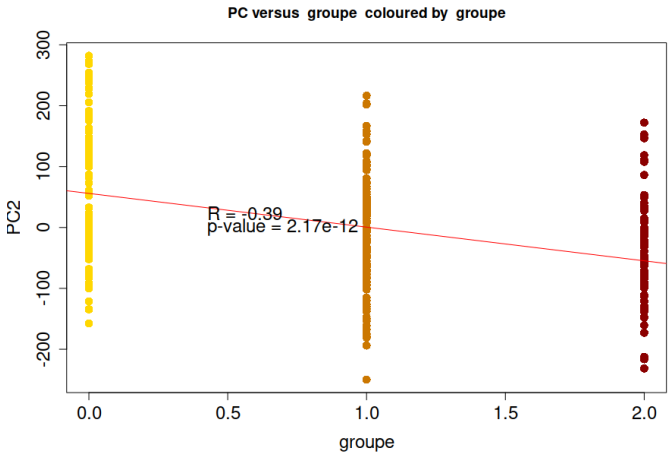
Coloured by age

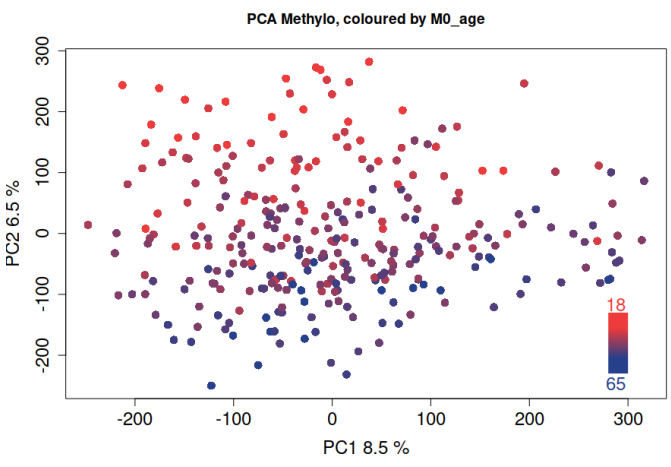


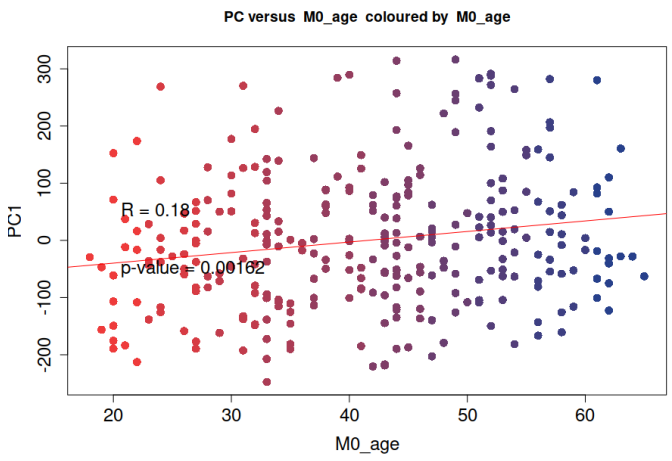
Correlation with PC1 and PC2

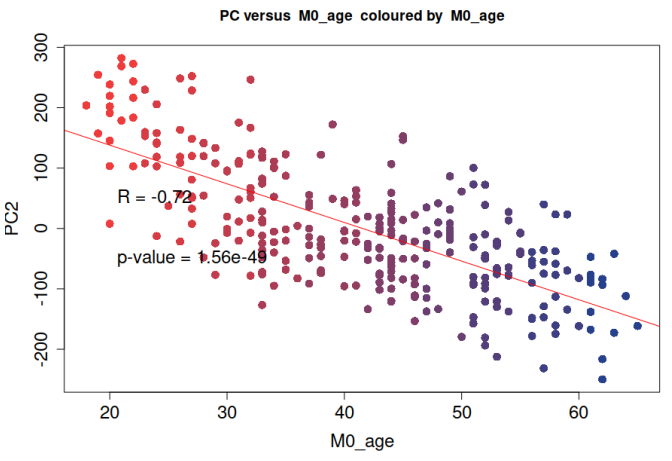












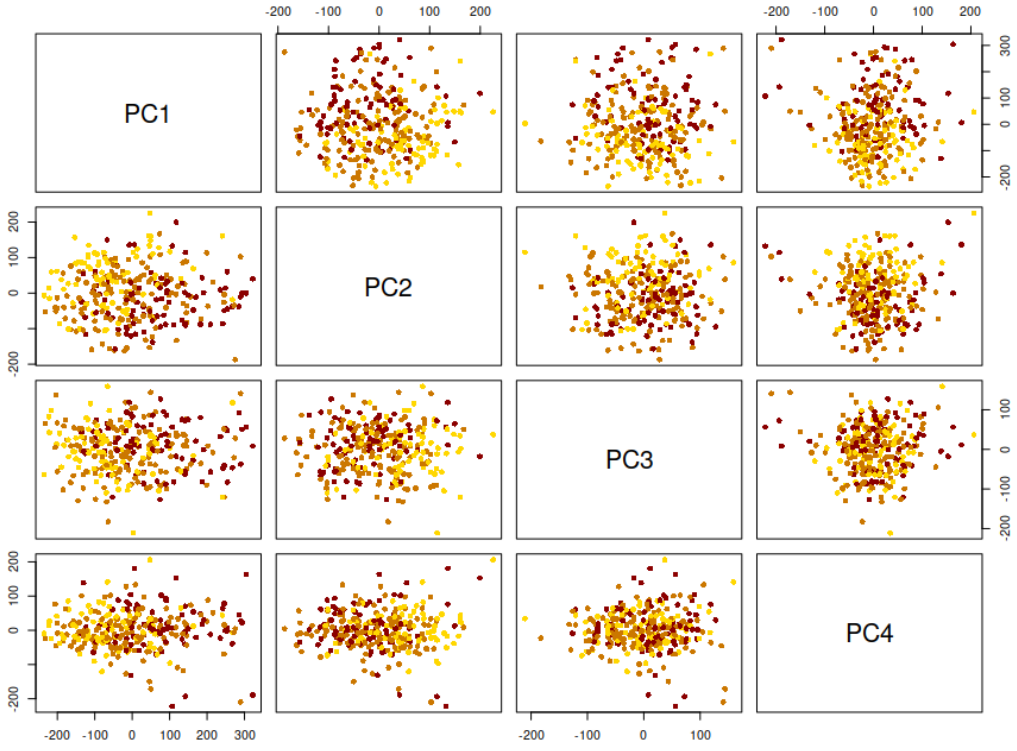
Removal of age component

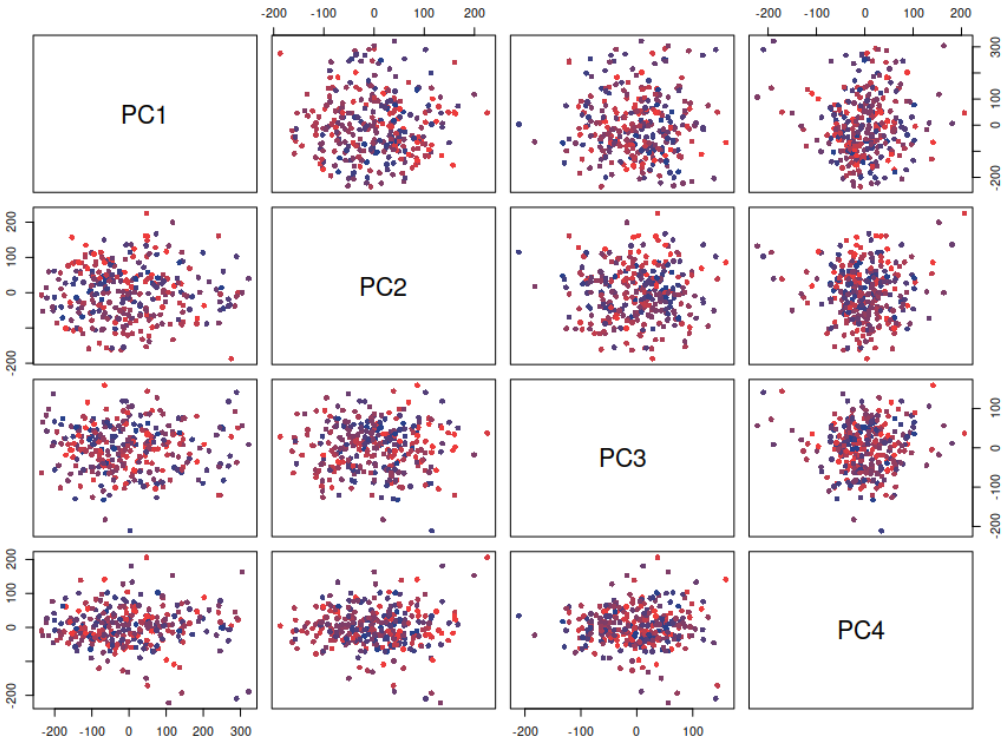
For each CpG i

$$\hat{M}_i = a_0 + a_1 \times age_i + \varepsilon_i$$

Keep the residuals

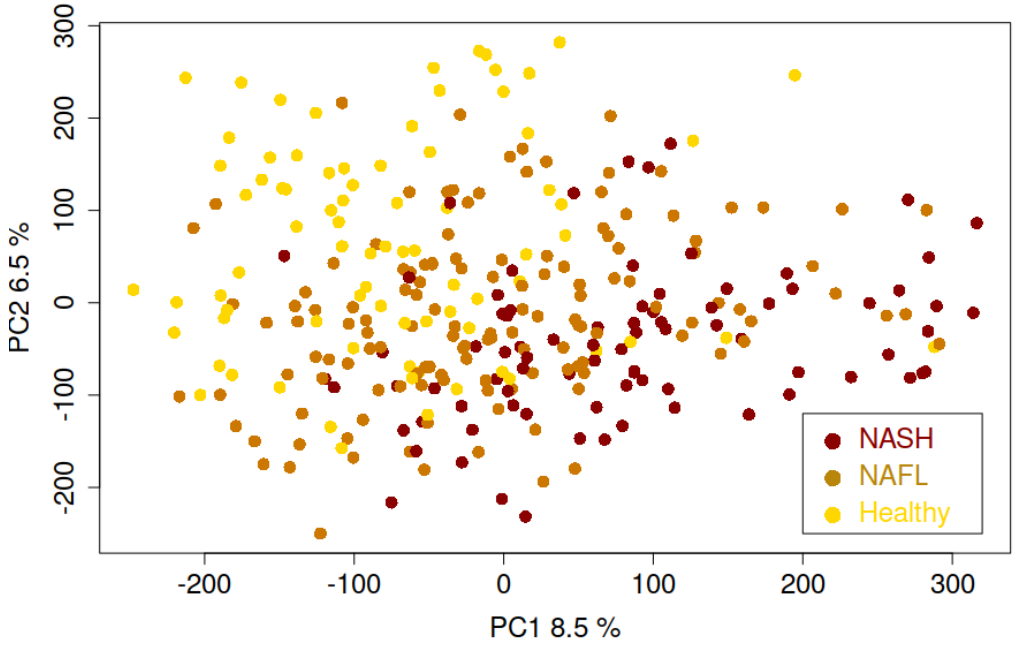
Bam! The age component is gone… but not (all of) the severity



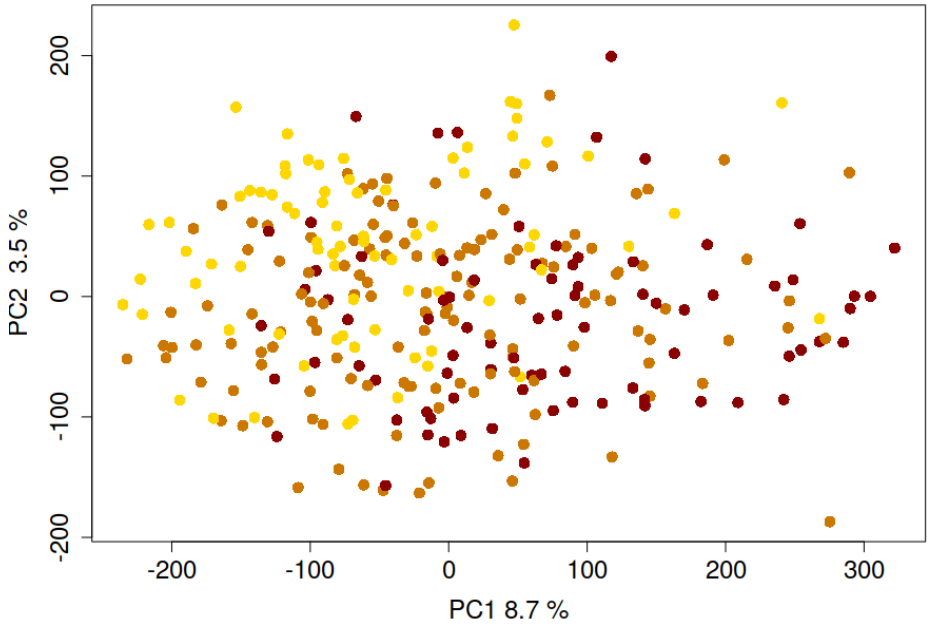


Bam! The age component is gone… but not (all of) the severity

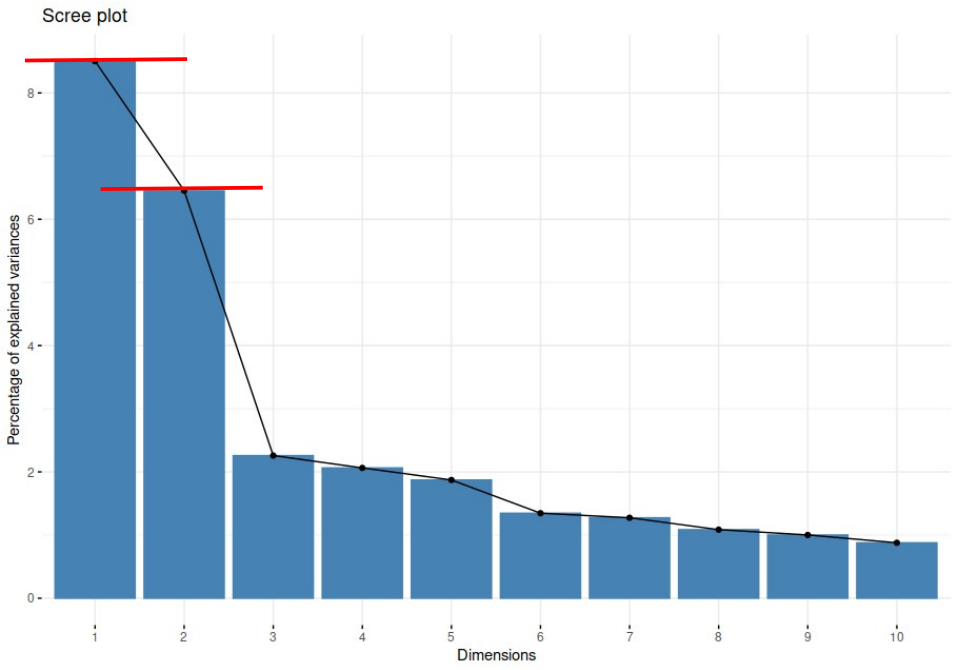
Before correction

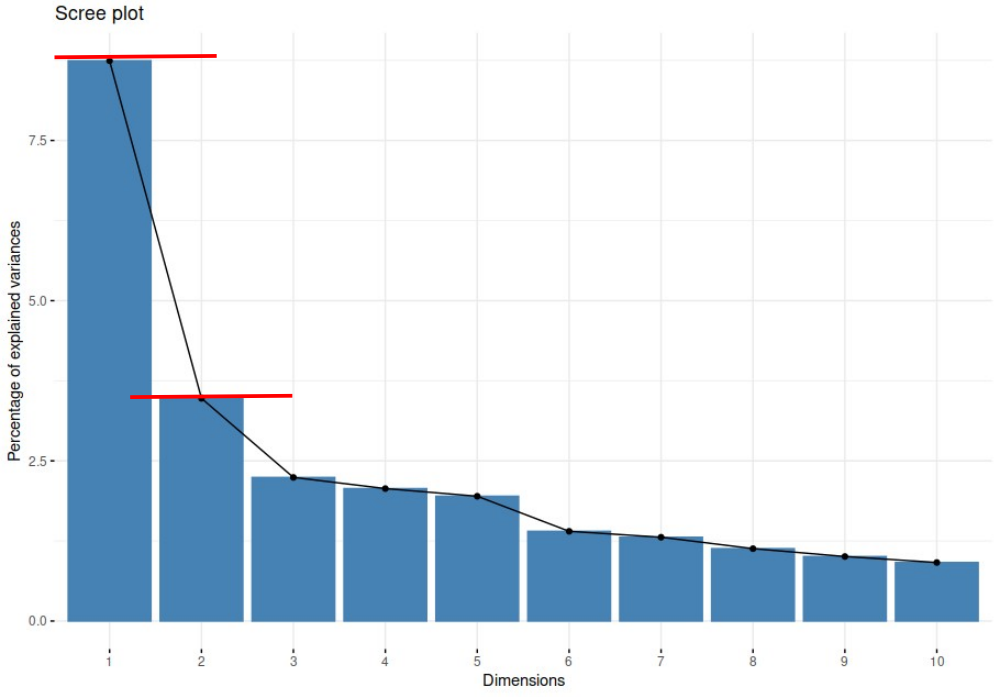


After correction



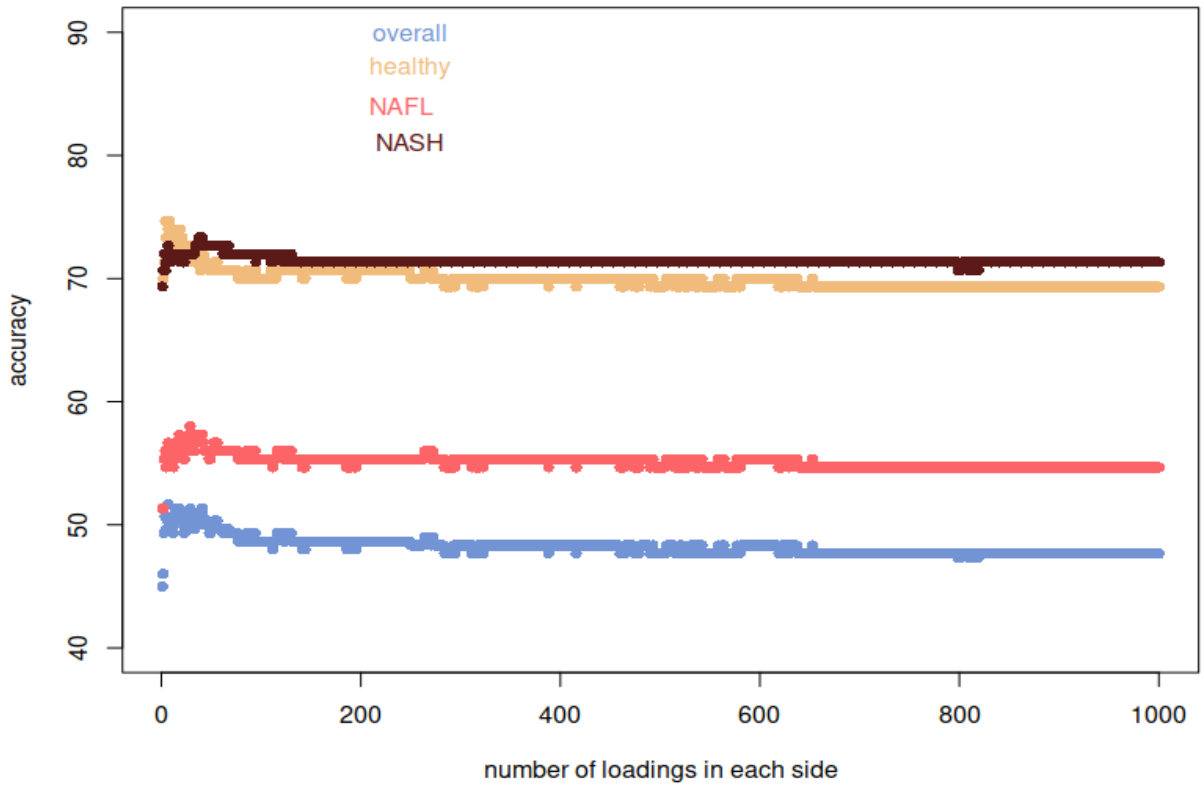
Transfer of variance from PC2 to PC1





Methyloscore

$$\sum_{i=1}^{\#\mathrm{CpGs}} \mathrm{Methyl}(\mathrm{CpG_i}) \times \mathrm{PC1load}(\mathrm{CpG_i})$$



Multi-Layer Perceptron trained on CpG methylation

130+130 CpGs (beta values!)

Dropout 30%

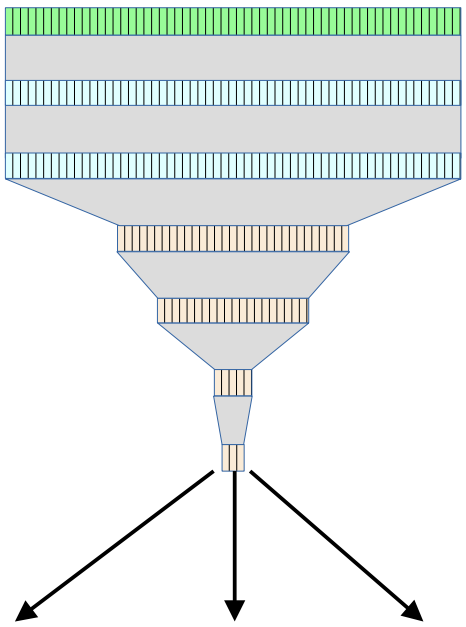
Normalisation

50 tanh

20 tanh

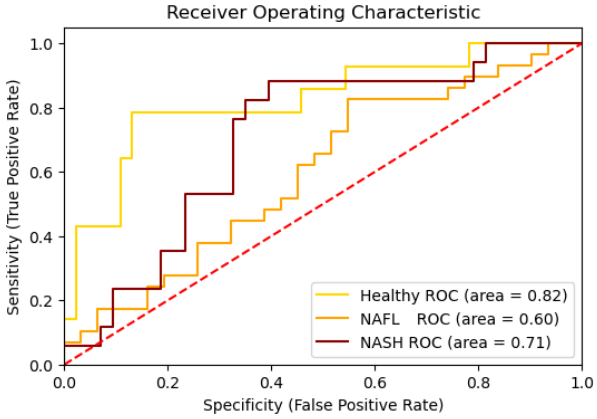
5 sigmoid

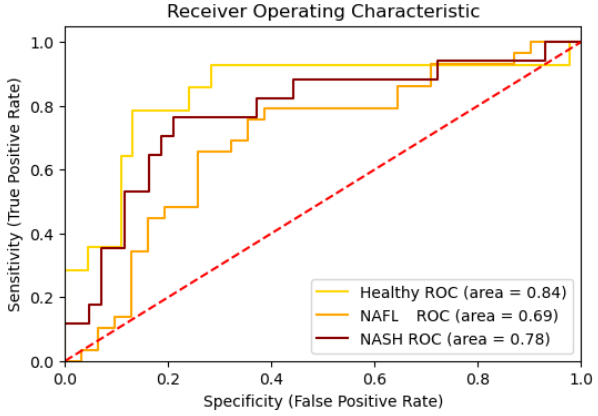
3 softmax

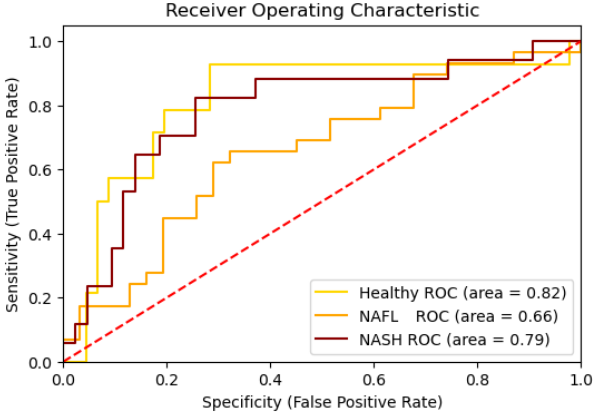


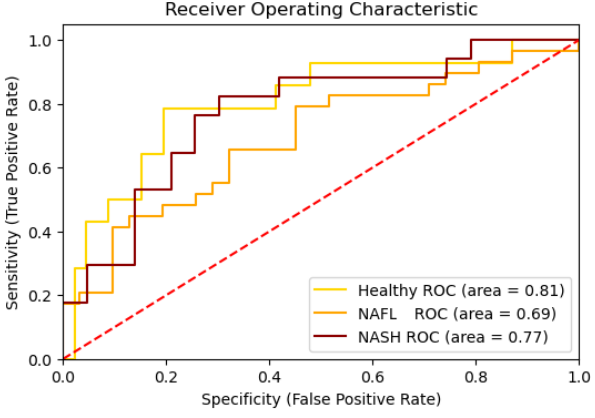
Healthy NAFL NASH

Methylation data predictivity on the independent dataset

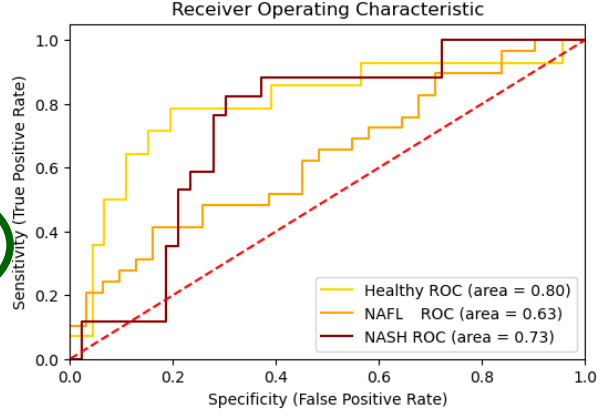






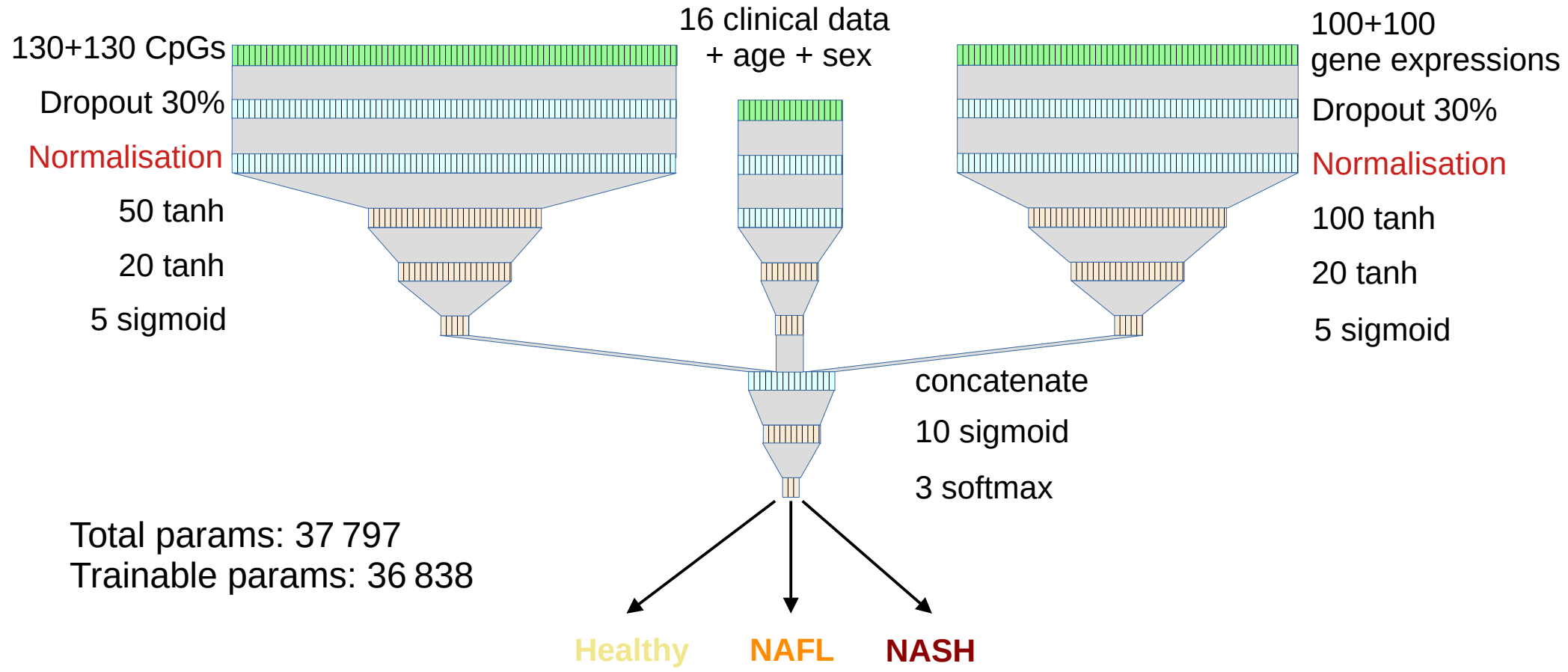


	healthy	NAFL	NASH	overall
TP	11	12	10	
TN	34	22	37	
FP	13	6	8	
FN	2	20	5	
AUC	0.78	0.58	0.74	0.67
accuracy	75.00	56.67	78.33	55.00
sensitivity	72.34	78.57	82.22	77.71
specificity	84.62	37.50	66.67	62.93
precision	45.83	66.67	55.56	56.02



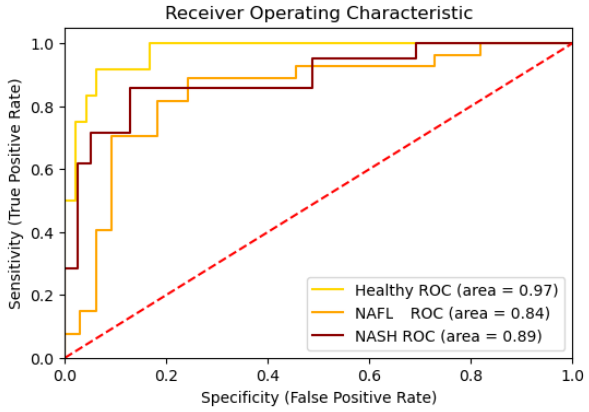
Supercharge predictions by putting everything together

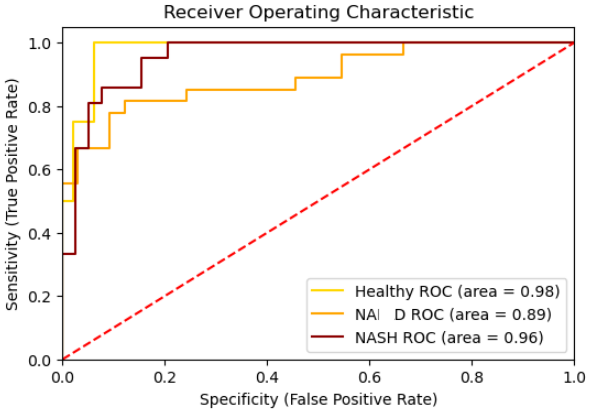
Methylation+ClinDat+RNAseq data

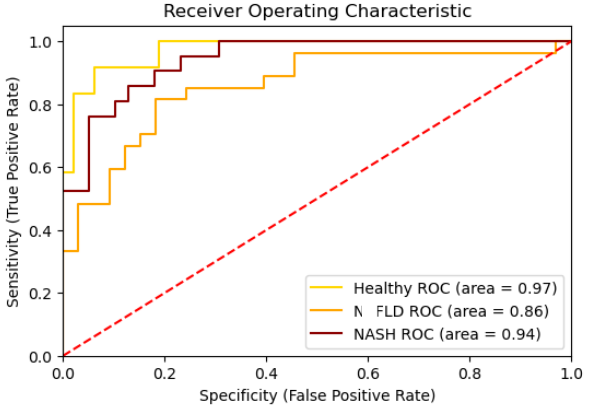


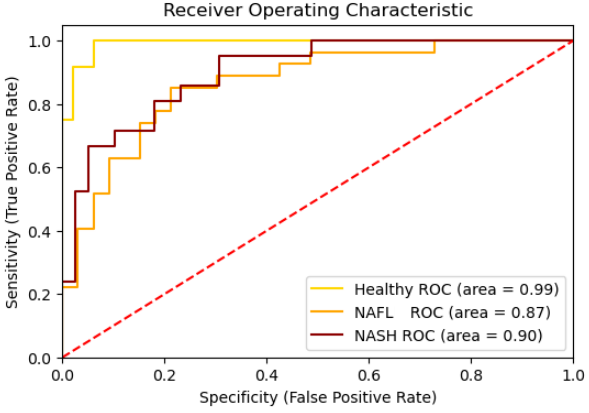
Total params: 37 797
Trainable params: 36 838

Predictivity of the MLP trained on 3 types of data

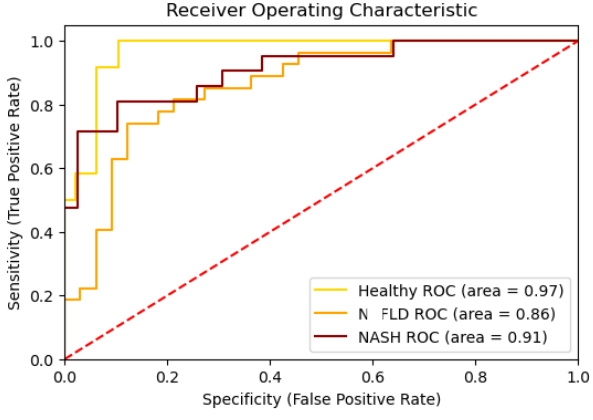








	healthy	NAFL	NASH	overall
TP	12	24	17	
TN	45	30	38	
FP	3	3	1	
FN	0	3	4	
AUC	0.97	0.90	0.89	0.91
accuracy	95.00	90.00	91.67	88.33
sensitivity	93.75	90.91	97.44	94.03
specificity	100.00	88.89	80.95	89.95
precision	80.00	88.89	94.44	87.78



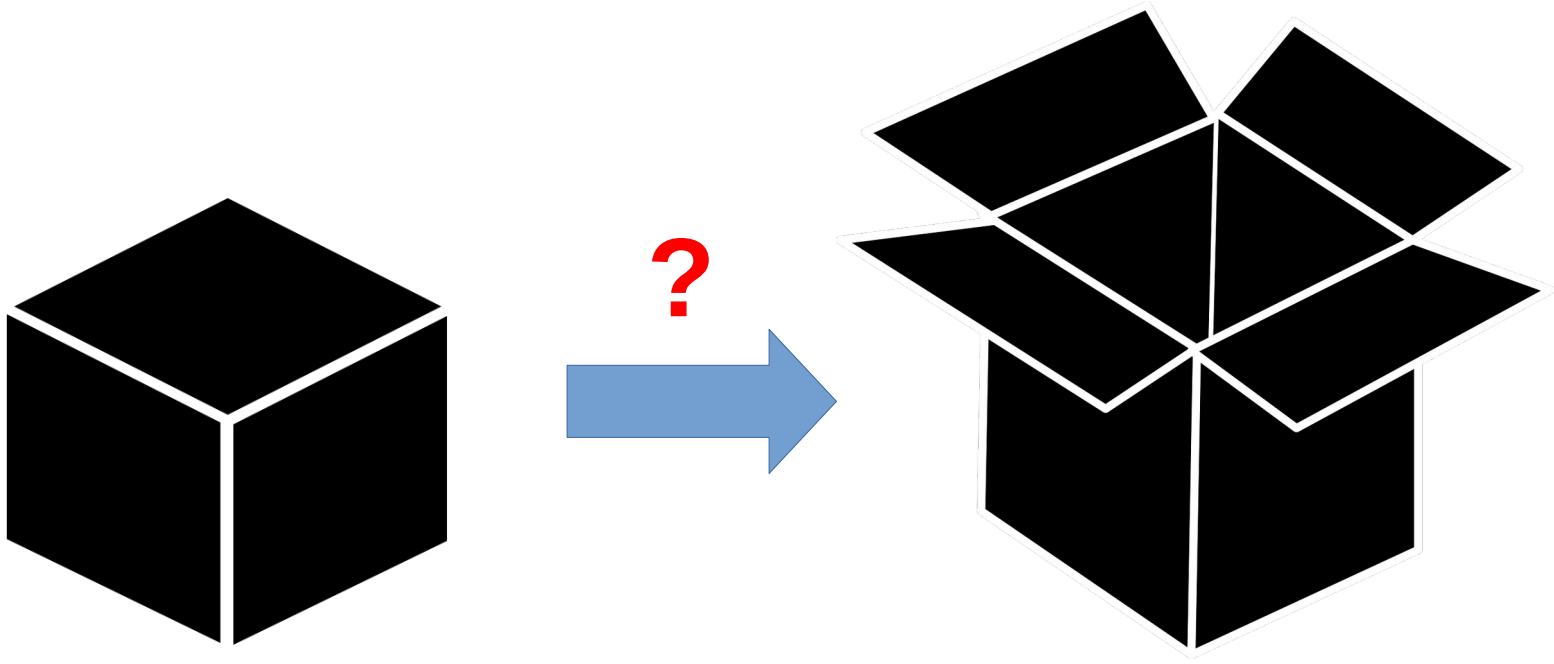
Comparison with some prior tools

Tanwar *et al* (2013). Procollagen III terminal peptide. AUC Healthy from NAFLD **0.88** (here **0.97**); AUC NASH from NAFL **0.79** (here **0.89**)

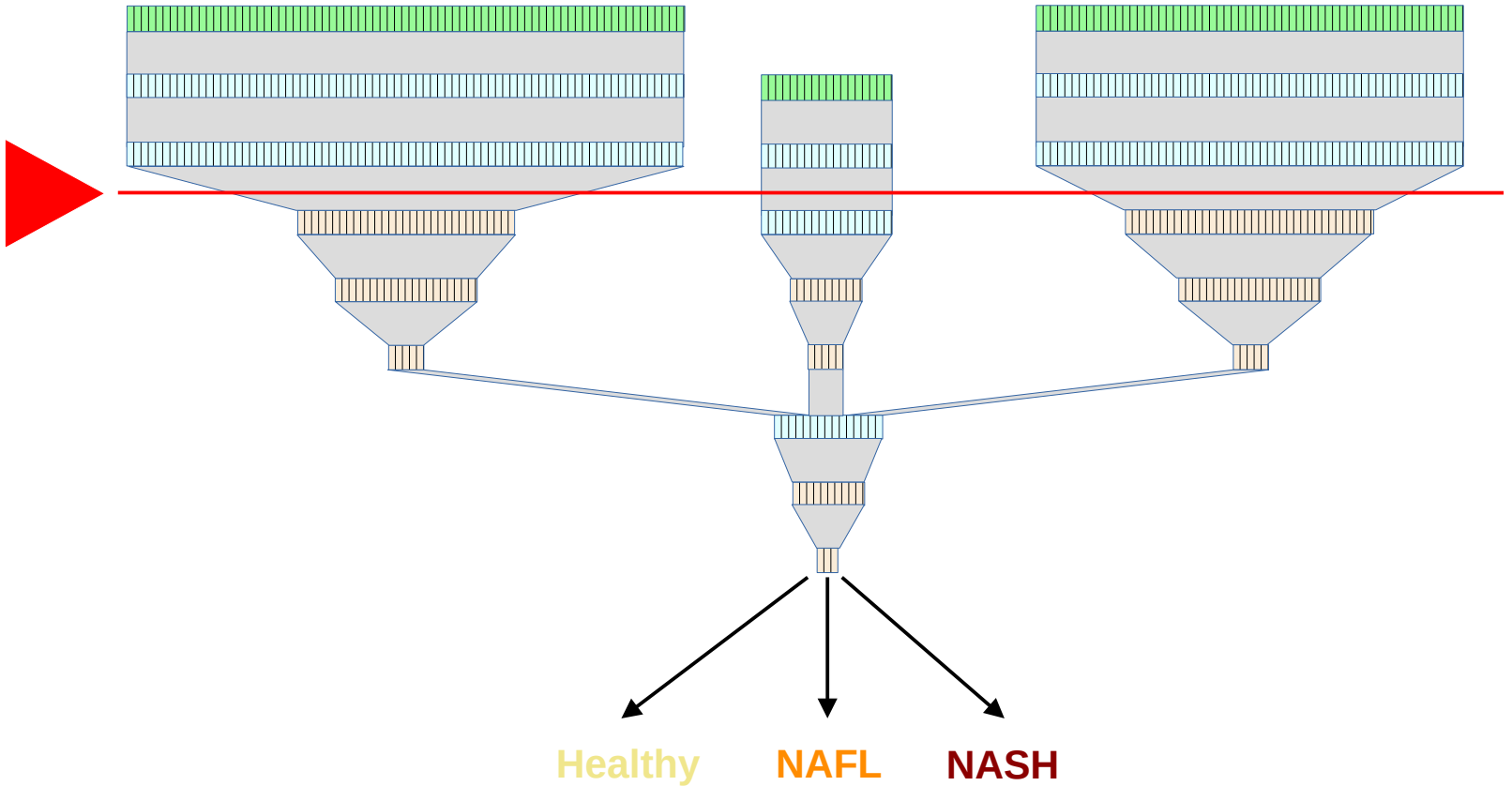
Sanyal *et al* (2018). miR-34a, α2-macroglobulin, YKL-40, HbA1c. AUC NASH from non-NASH 0.81 (here 0.89), accuracy of **72%** (here **91.67%**)

Park *et al* (2023). Carefully hand-picked 10 genes. Only distinguish NASH from NAFL. Accuracy of NASH prediction on an independent cohort is **80.5%** (here **91.67%**)
6 of their genes are not in our set
(CDC6, TTK, HASPIN, MCM10, SLC7A1, MAD2L1)

Can we explain how a model decides?



Importance of inputs

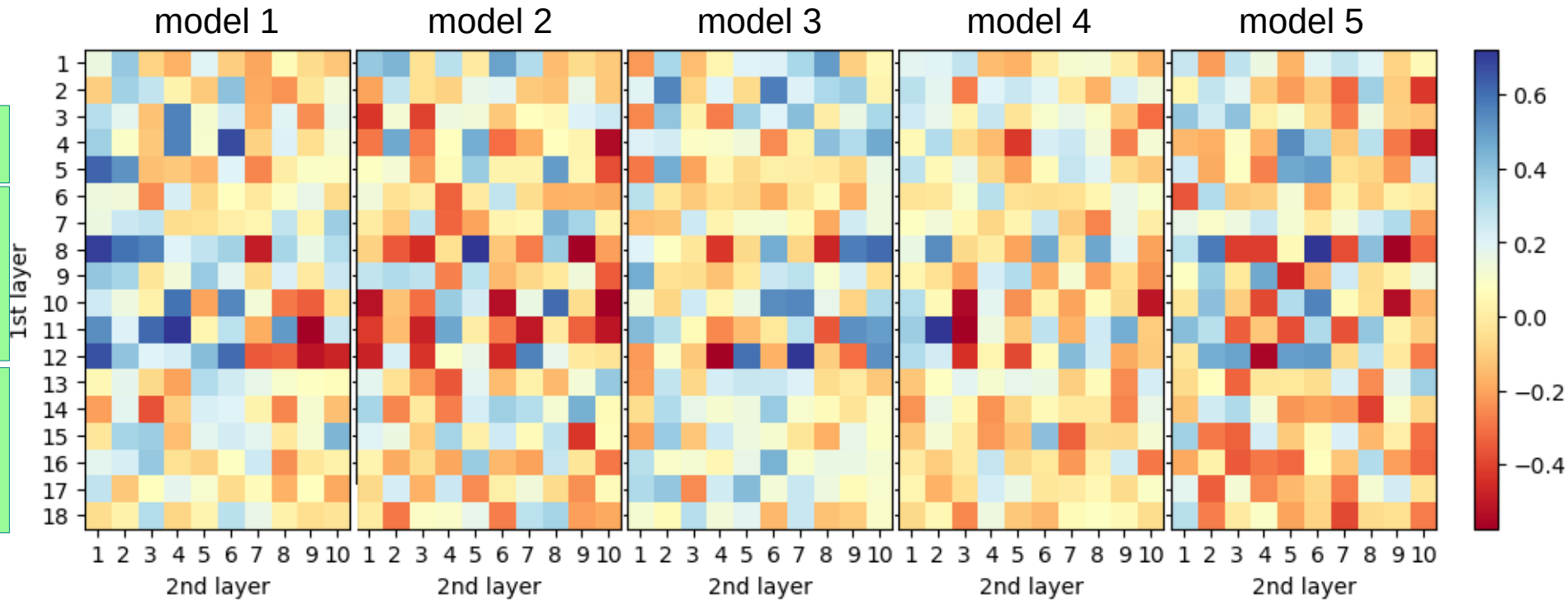


Clinical data module weights, inputs

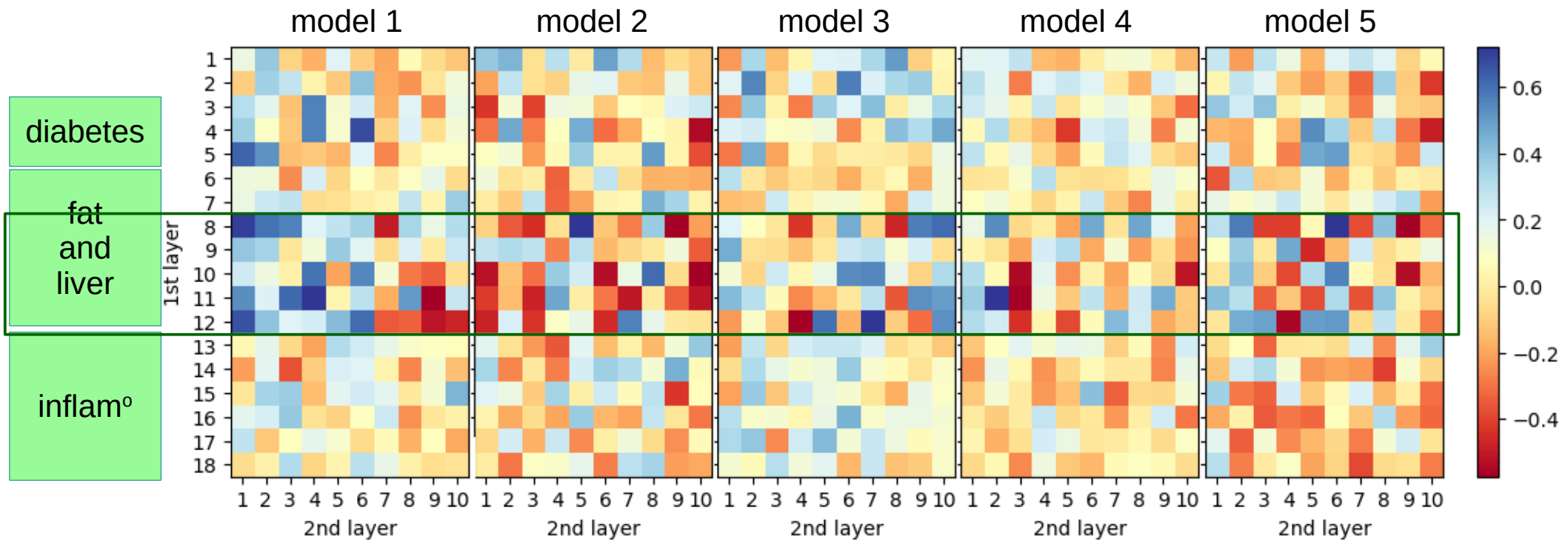
diabetes

fat and liver

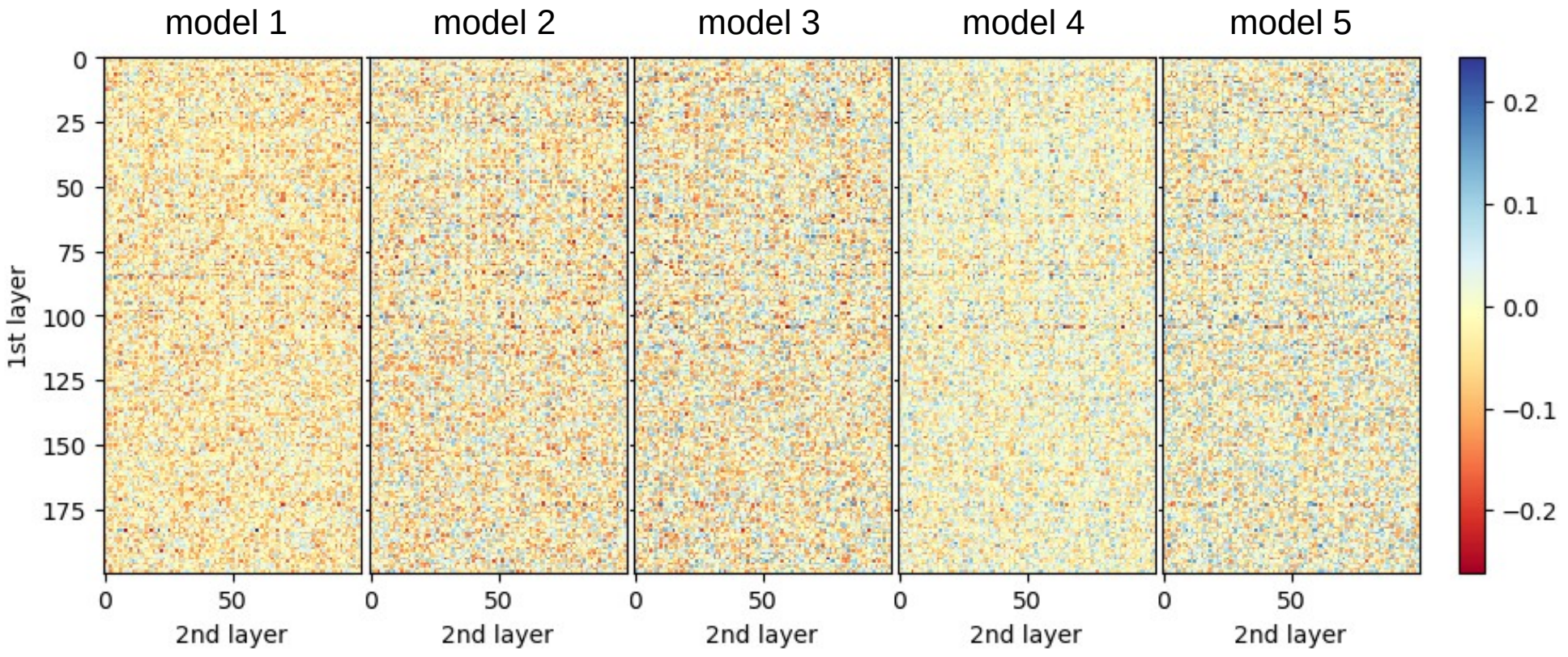
inflammation



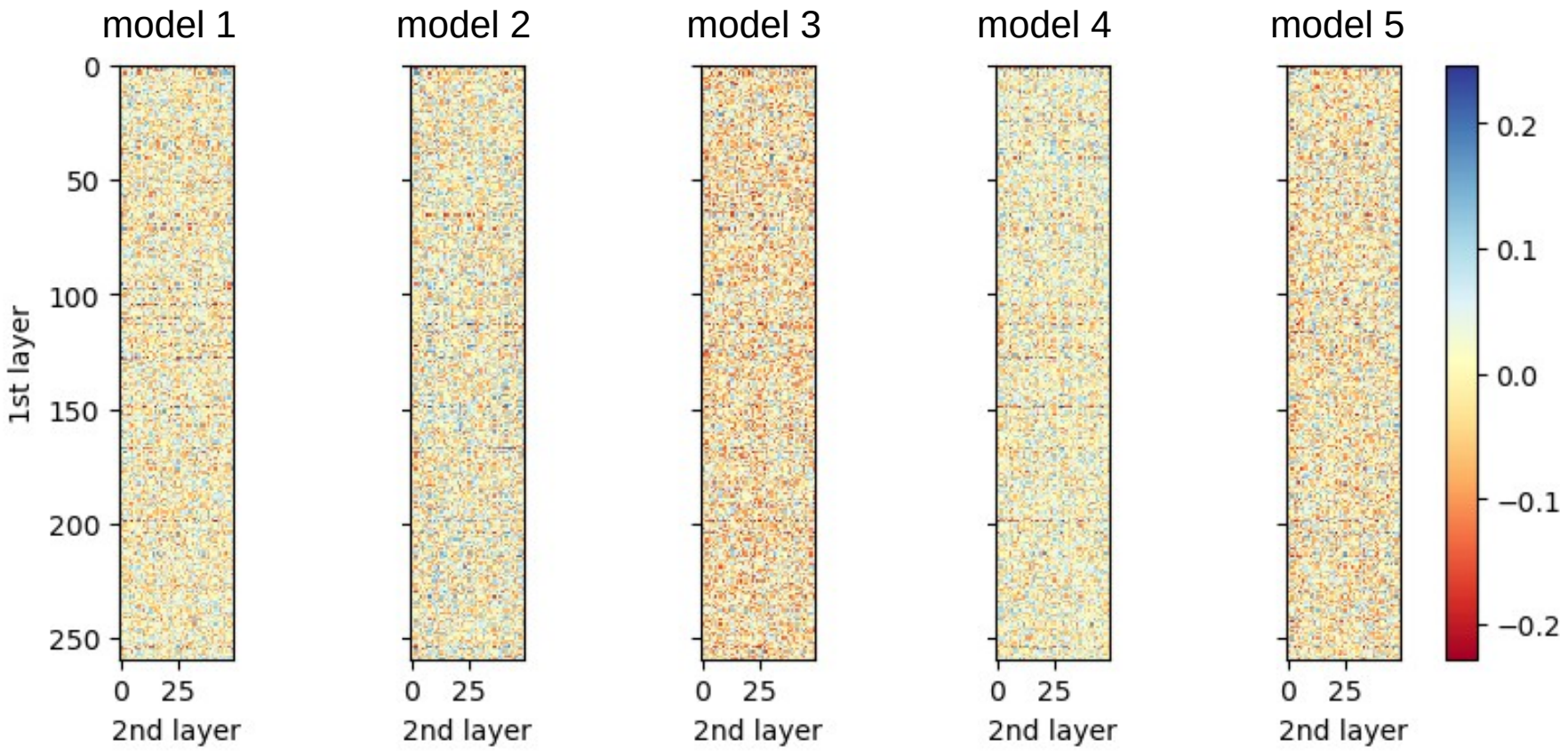
Clinical data module weights, inputs



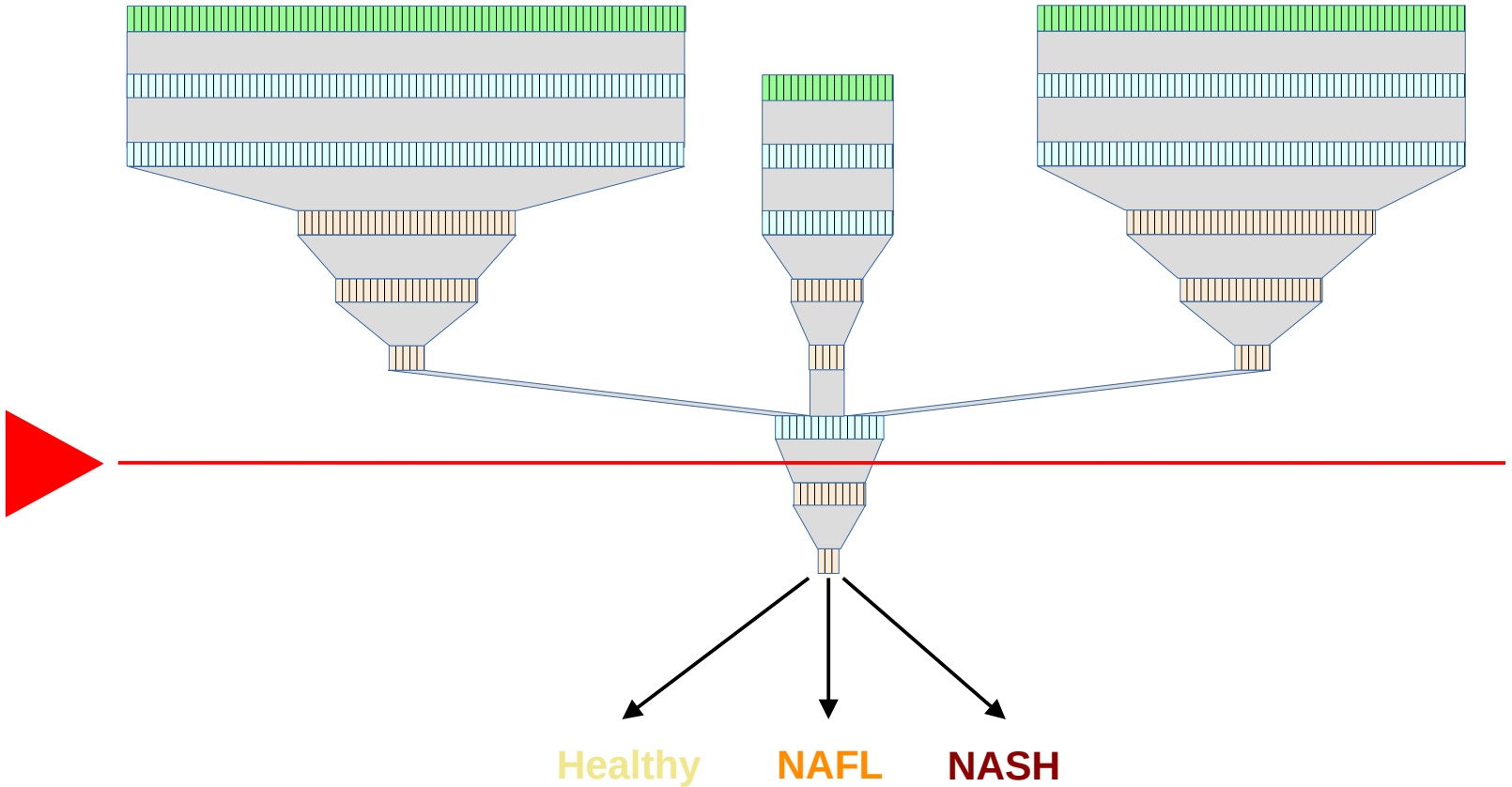
RNAseq module weights, inputs



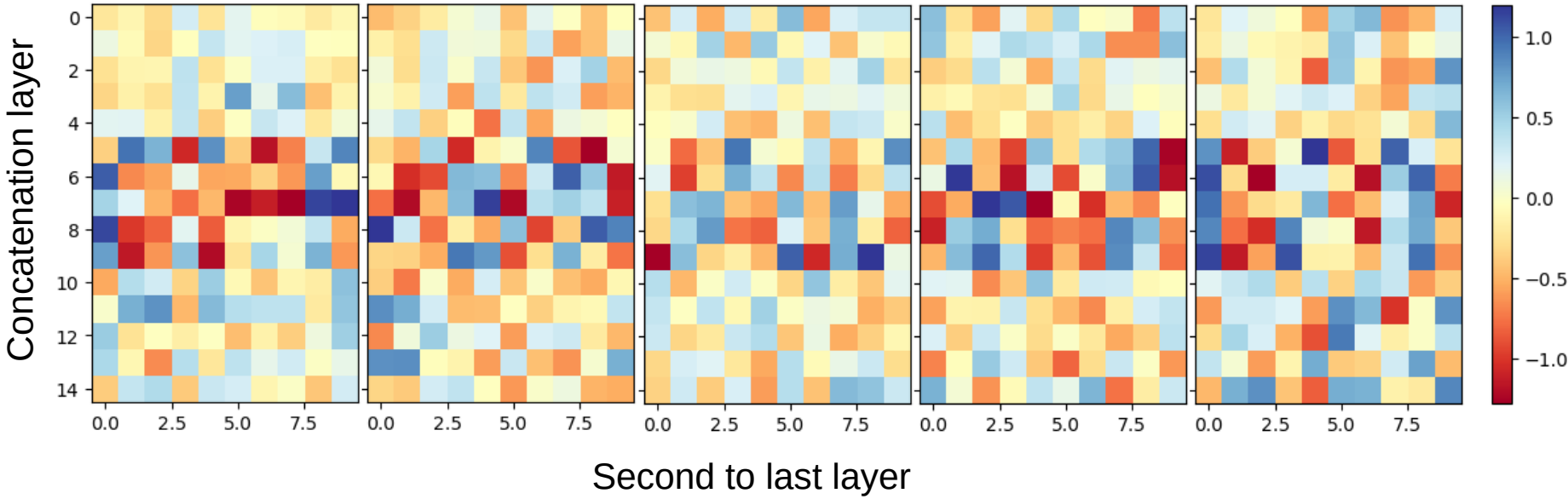
Methylation module weights, inputs



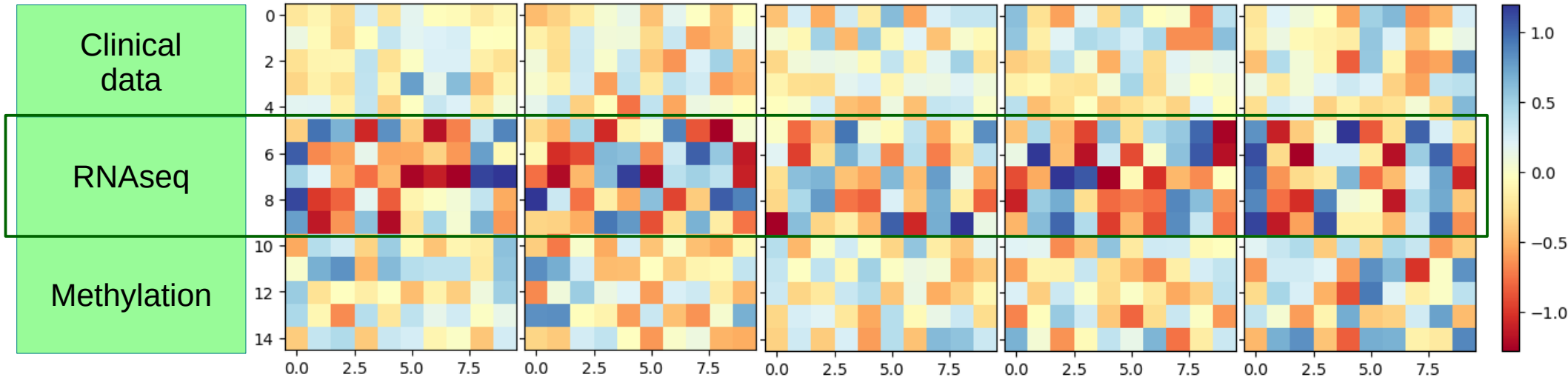
Importance of modules



Impact of the different modules after concatenation

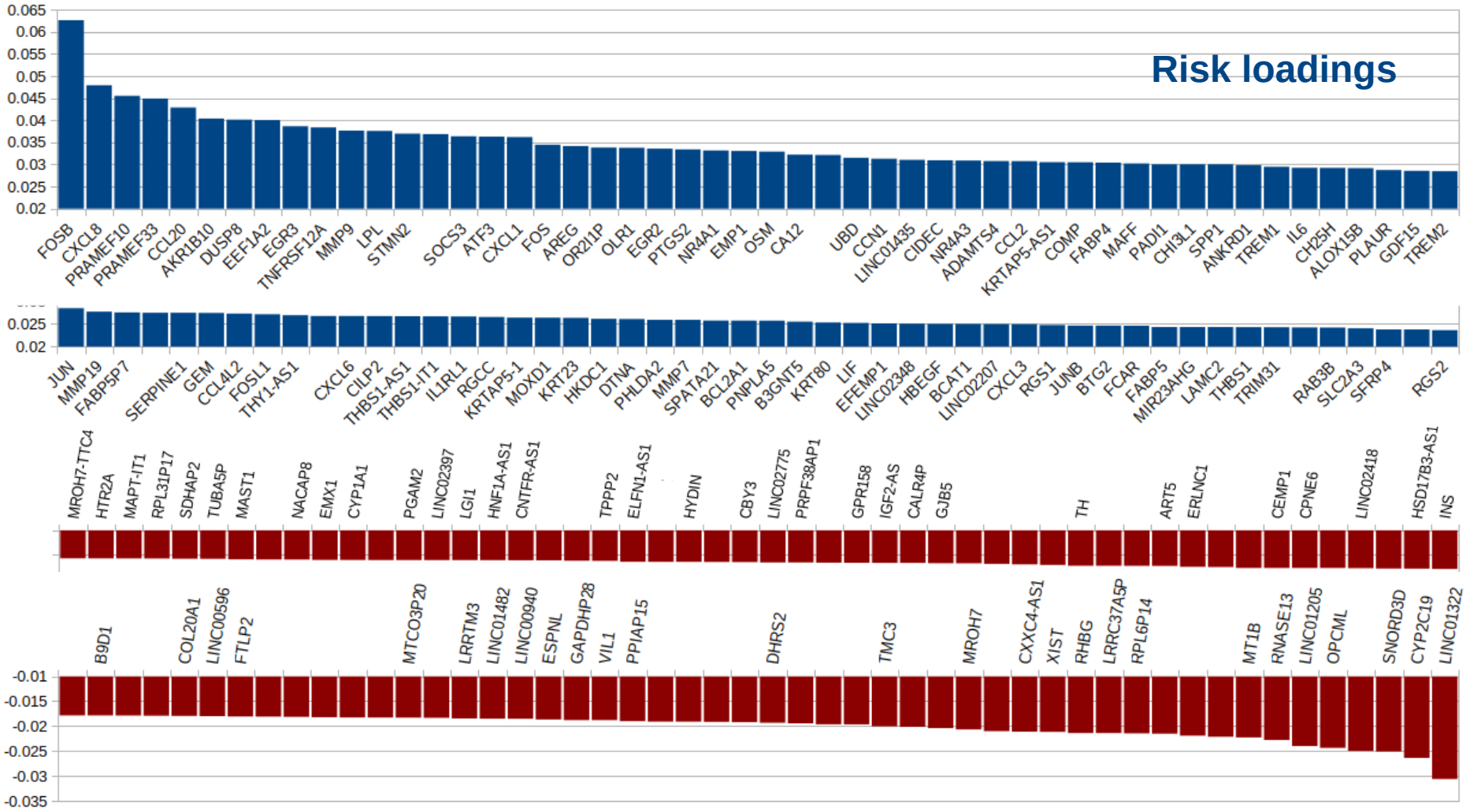


Impact of the different modules after concatenation



In all five models, the RNAseq module has the main impact

100 protection and risk loadings on PC3



What now? This project

Making the tools more useful:

We need to incorporate metabolomics and genotypes.
The latter will require more complex architectures
(one-hot encodings+CNN, genome-informed local structures, etc.)

We need tools using blood molecular phenotypes as input:
1) non-invasive, 2) with a liver biopsy, we do not need molecular phenotypes…

Understanding the basis of the decisions, and the relationships between inputs:

What happened to PreciNASH subjects? NASH FP became TP? HCC?

Analyse the data generated during model training
(loadings for RNAseq and methylation, age-corrected methylation, etc.)

Interpretable AI approaches, including perturbation (e.g., ablation)
and addition of attention structures.

What now? EGID and PreciDIAB

Nobody can ignore AI; Nobody should ignore AI

AI is part of a PreciDIAB workpackage (and of almost any recent project currently ongoing or submitted by UMR 1283/8199)
☛ AI is becoming one of the core tools of UMR 1283/8199

We need more deep learning approaches to learn from different data types (genotypes, ATAC-seq, chromosome capture, socioeconomic and psychological data, etc.):
e.g., prior knowledge embeddings, LLMs,
multi-omics attention, Vision Transformers

AI as a tool for discovery and explanation,
on par with estimation, prediction, and clustering

This requires scaling up → money, but more importantly brains.

Acknowledgements

Philippe Froguel
Amélie Bonnefond

Mathilde Boissel
Lijiao Ning
Emma Henriques
ShuangShuang Geng

Anne-Sophie Ledoux
Vincent Massy

Amna Khamis
Stefan Gaget
Mehdi Derhourhi

Hélène de Gavre
Mélanie Hocquet

François Pattou (PreciNASH)

R package developers
(*VIM*, *MICE*, *DESEq2*, *ChAMP*)

Python package developers
(*Tensorflow/Keras*, *Numpy*, *Pandas*, *Scikit-learn*)

Marie Le Novère

PreciDIAB
National Center for
Precision Diabetic Medicine