# Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women

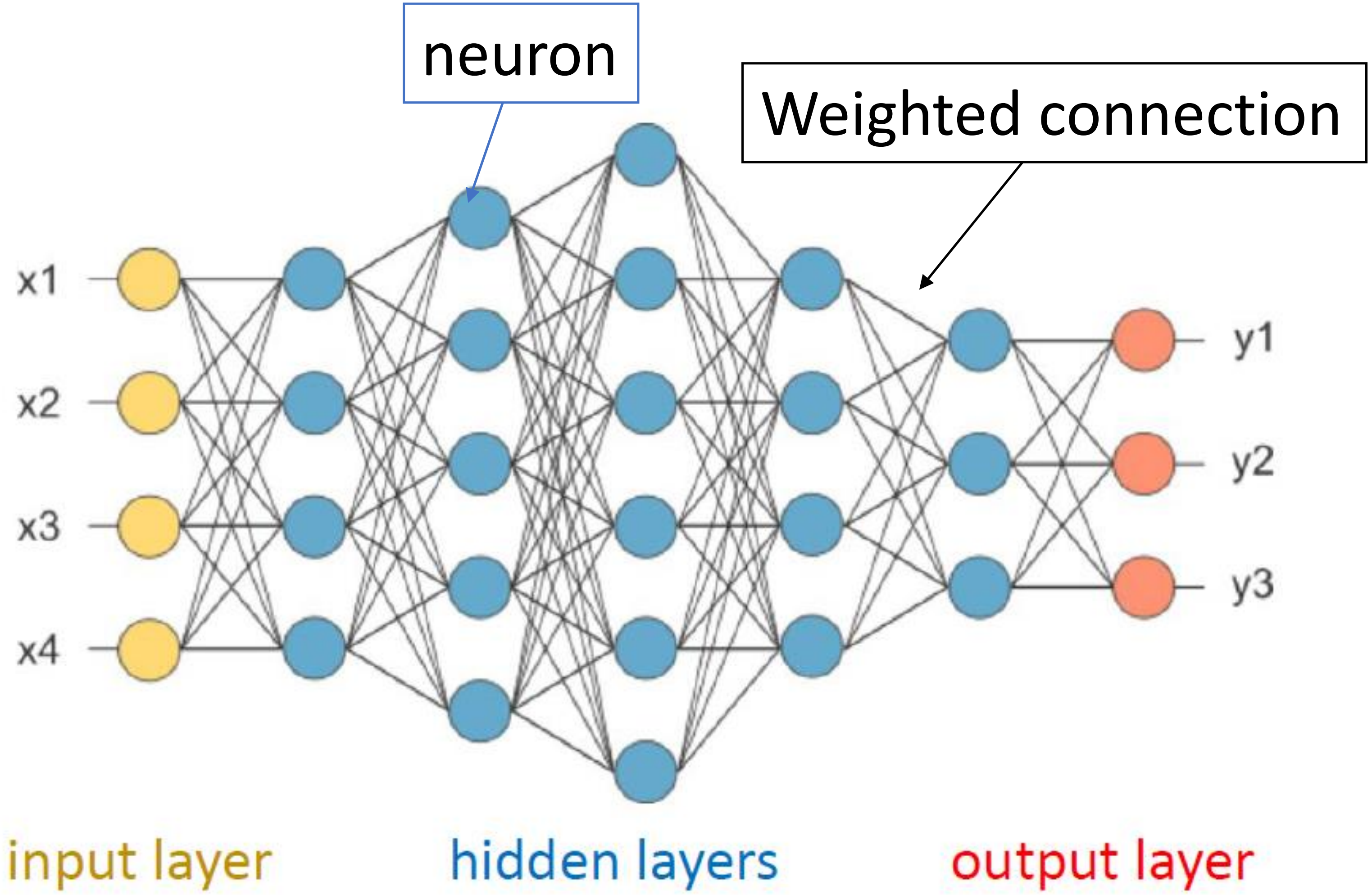
Paul Fergus, Casimiro Curbelo Montañez, Basma Abdulaimma, Paulo Lisboa, and Carl Chalmers

Presented by Ming Wai Yeung
10-04-2018

# Outline

- Brief introduction on deep learning
- The study
  - Motivation
  - Reason for using deep learning
  - Aim of study
  - Study design
  - Data and method
  - Result
  - Significance and comments

# A neural network



# Training a neural network

- Iteratively pass data through the network
- Assess the output, change the weights (edges) based on error
  - Forward →
  - Backward ⇢
  - Update parameters →
  - Gradient δ

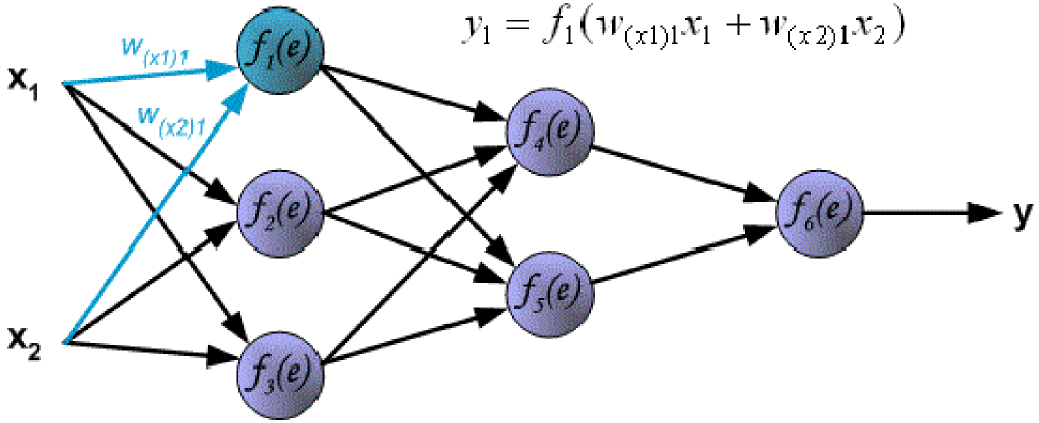
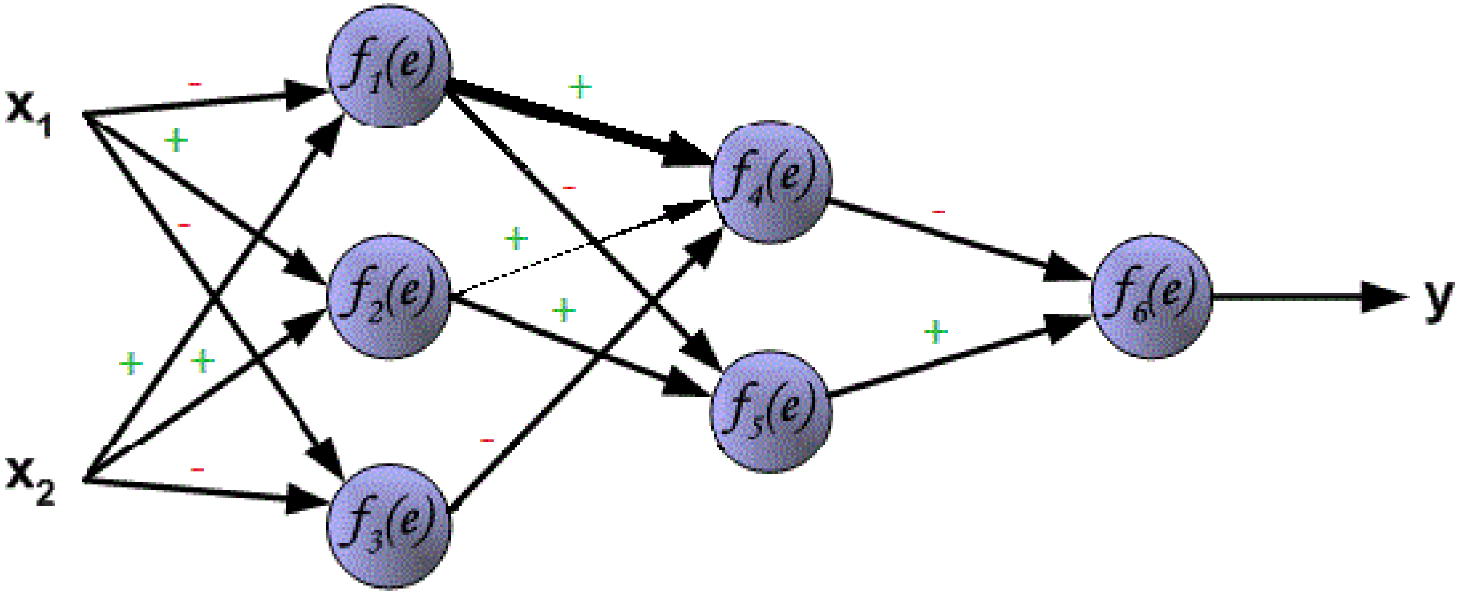


Image credit: http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

# A trained network



- Usage depends on structure of output layer
- For binary classification:
  - Output layer with 2 neurons
  - Each outputs the probability of belonging to one class

# Outline

- Brief introduction on deep learning
- The study
  - Motivation
  - Reason for using deep learning
  - Aim of study
  - Study design
  - Data and method
  - Result
  - Significance and comments

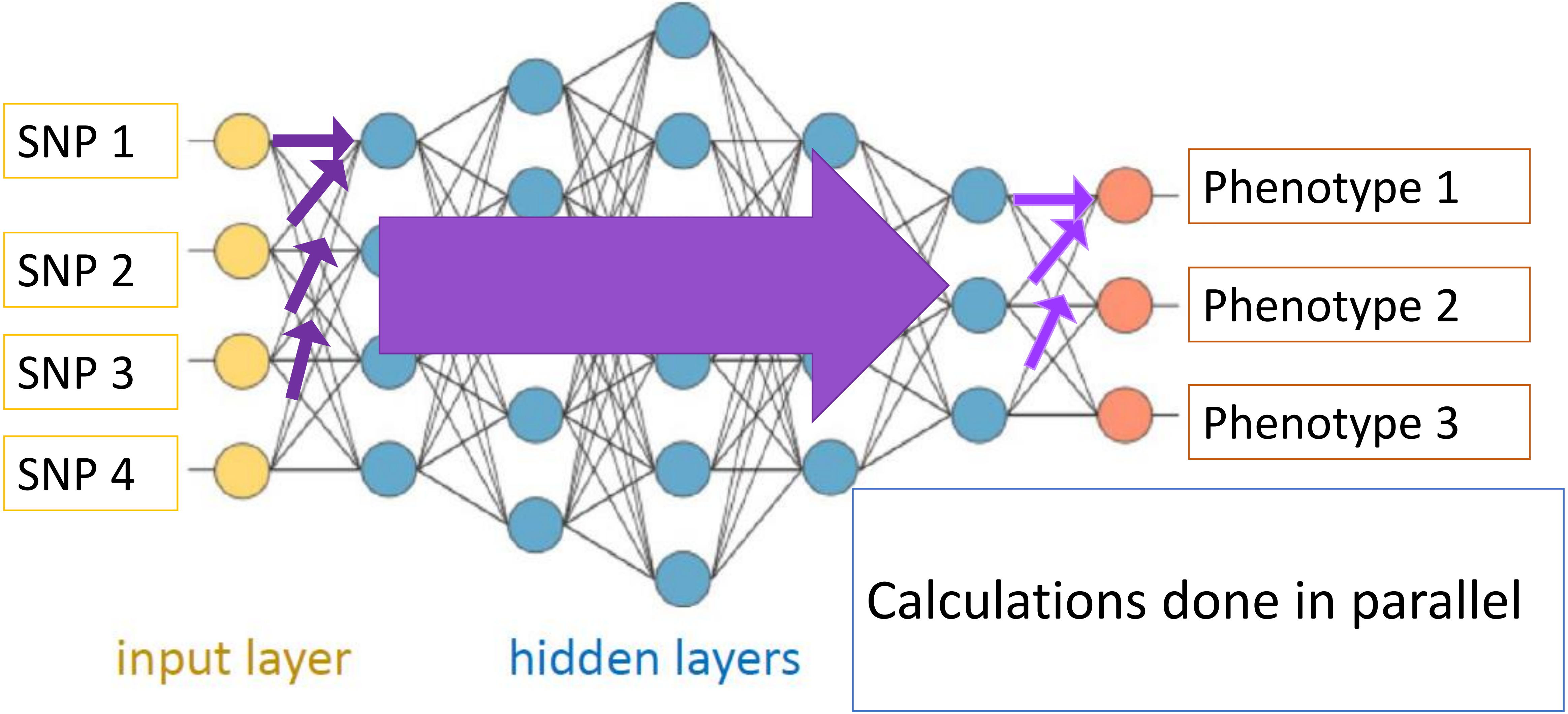
# Motivation

- Preterm Birth
    - Genetic component
        - 20-40% heritability
- Traditional GWAS looks at the SNPs individually
    - Epistasis - interactions between SNPs ignored

    → PLINK Epistasis tests

        ! Scalability issues still persist

# Capture the interaction with neural network



Calculations done in parallel

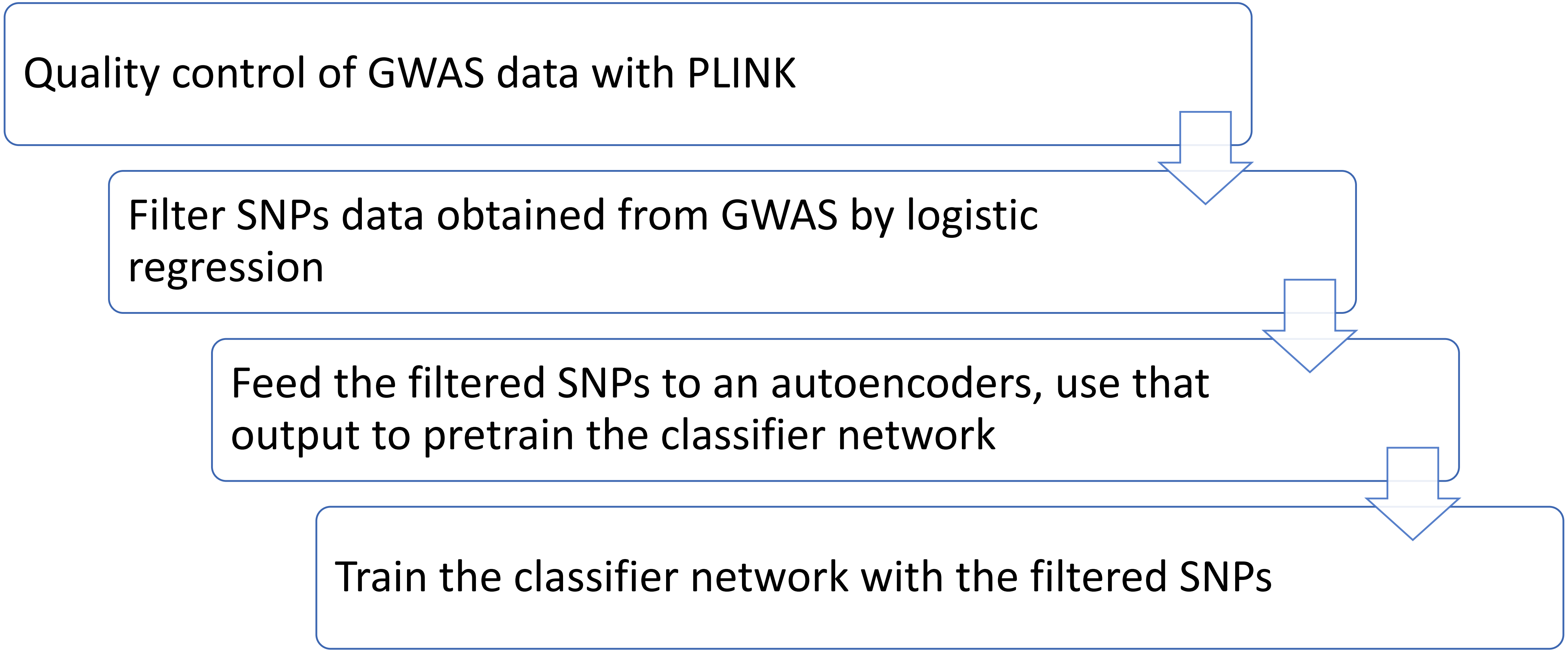
# Outline

- Brief introduction on deep learning
- The study
  - Motivation
  - Reason for using deep learning
  - Aim of study
  - Study design
  - Data and method
  - Result
  - Significance and comments

# Aim of study

- Present a deep learning framework in GWAS analysis
  - Extract latent representations capturing epistatic effects of major and minor SNP perturbations from GWAS data using a stacked autoencoder
  - Train a neural network to classify preterm birth

# Study Design



Quality control of GWAS data with PLINK

Filter SNPs data obtained from GWAS by logistic regression

Feed the filtered SNPs to an autoencoders, use that output to pretrain the classifier network

Train the classifier network with the filtered SNPs

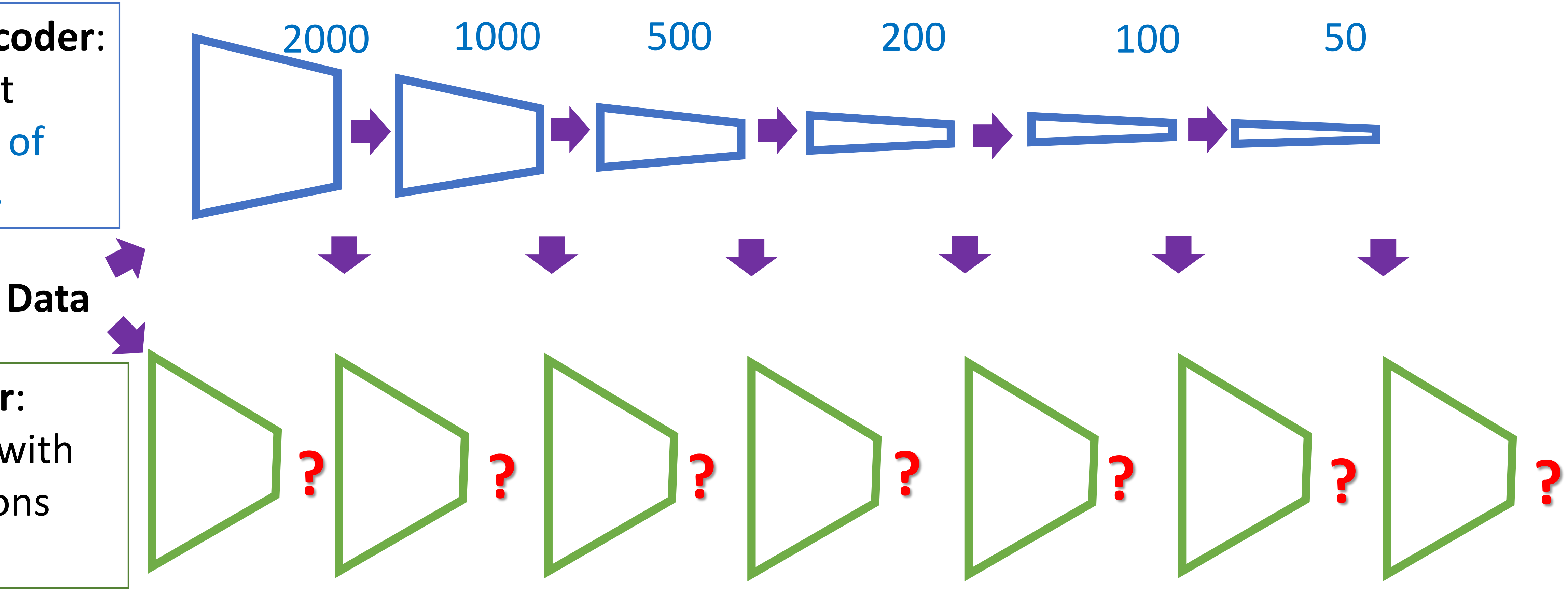
# Data and method

- 1000 case (mothers who delivered preterm) vs. 1000 age-matched control genotyped
    - After QC: 632 case , 895 control
- 2,362,044 SNPs for each individual
    - After QC:  1,927,820

# Deep learning

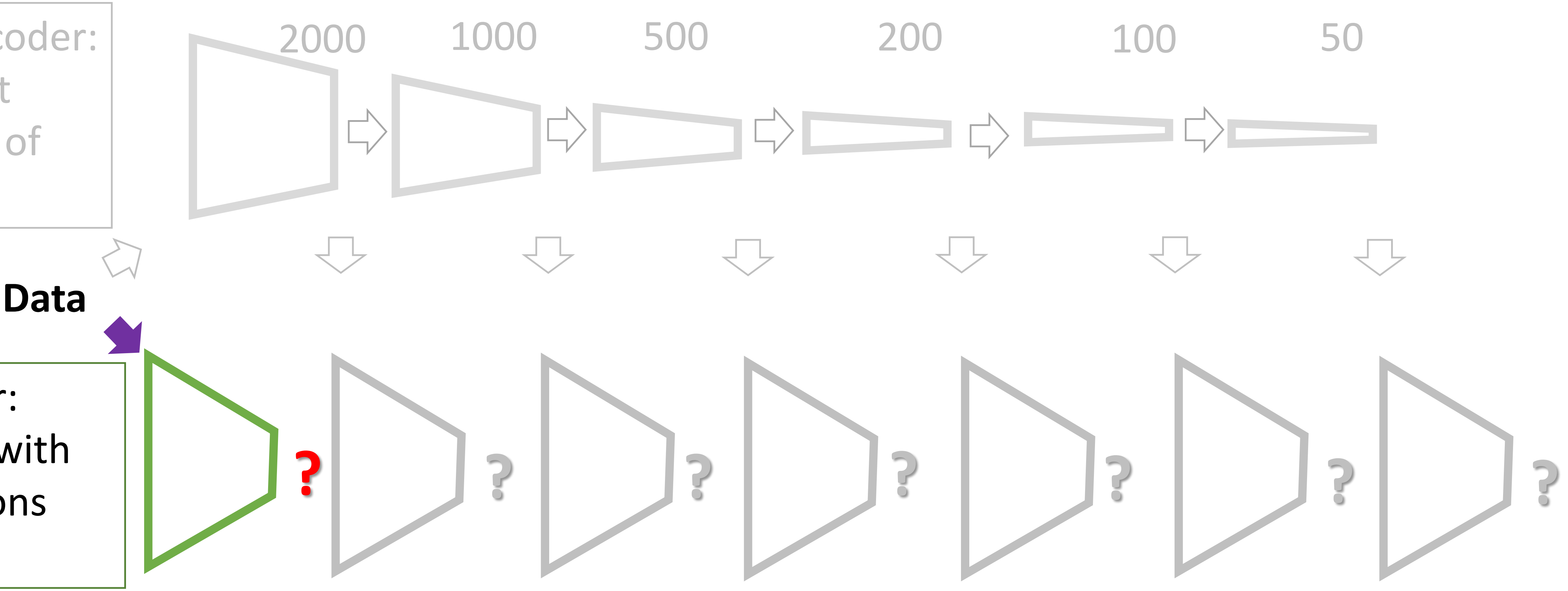
**Autoencoder**: Different number of neurons

**Classifier**: 4 layers with 10 neurons each



# Result

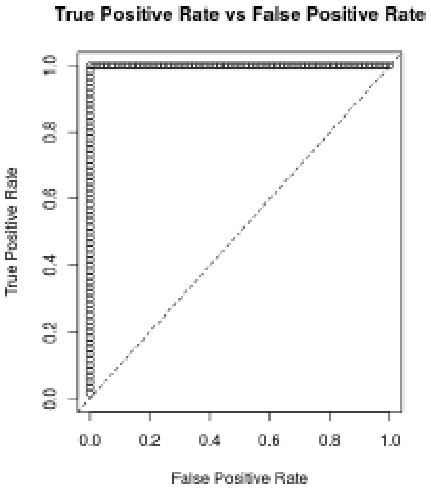
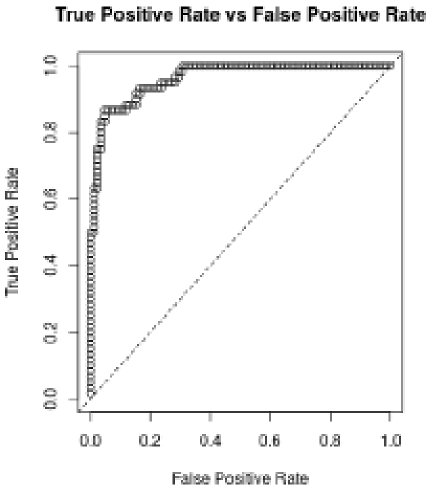
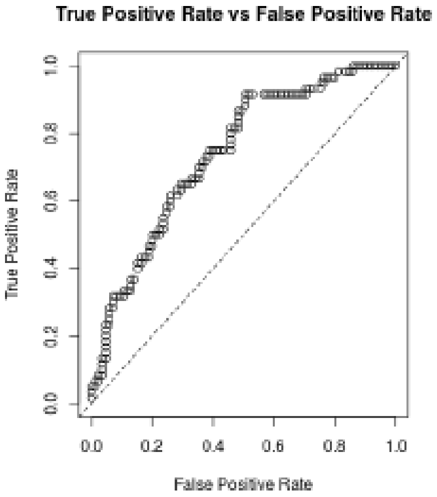
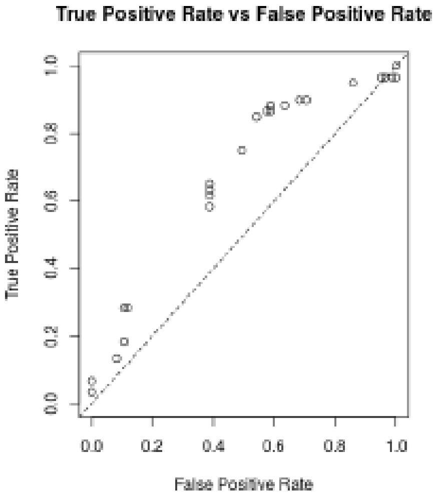
Autoencoder: Different number of neurons



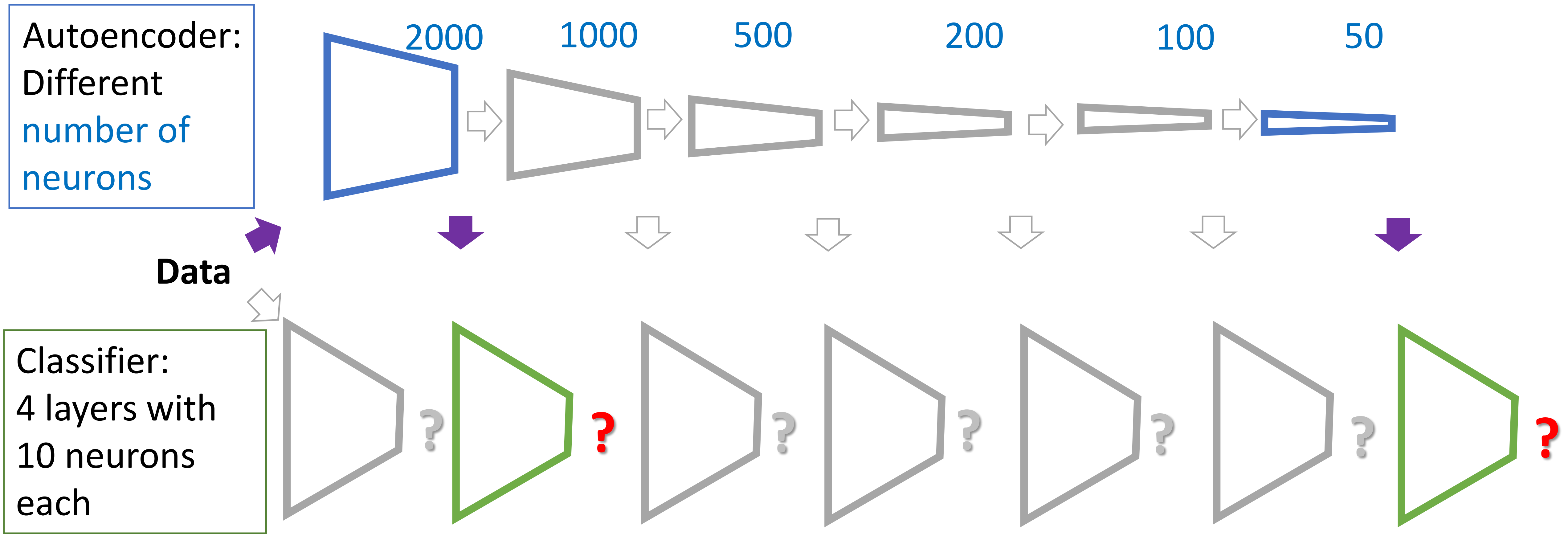
Classifier: 4 layers with 10 neurons each

# Performance on test set

- Without processing by autoencoder
- SNPs filtered by logistic regression with different thresholds

| | | | | |
|---|---|---|---|---|
| Number of SNPs as input | 4666 | 419 | 11 | **3** |
| |  |  |  |  |
| | (a) ROC for $5 * 10^{-3}$ | (b) ROC for $5 * 10^{-4}$ | (e) ROC for $5 * 10^{-7}$ | (f) ROC for $5 * 10^{-8}$ |
| AUC<br>(Area under curve) | **0.9998** | 0.9694 | 0.7416 | 0.6786 |

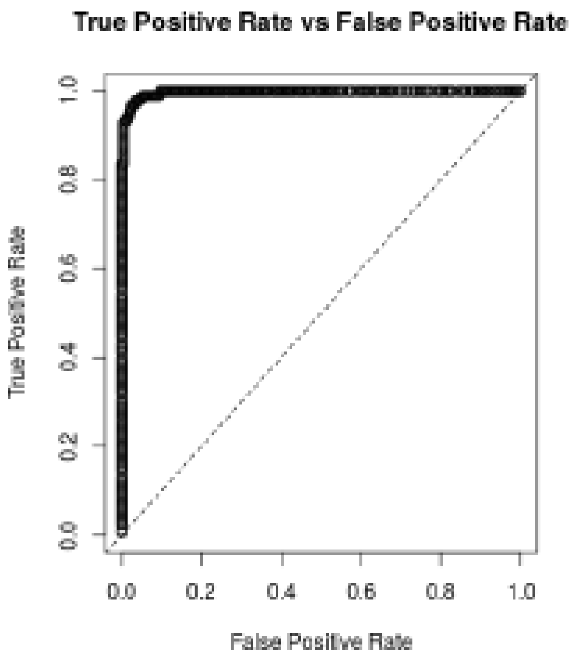
# With processing by autoencoder



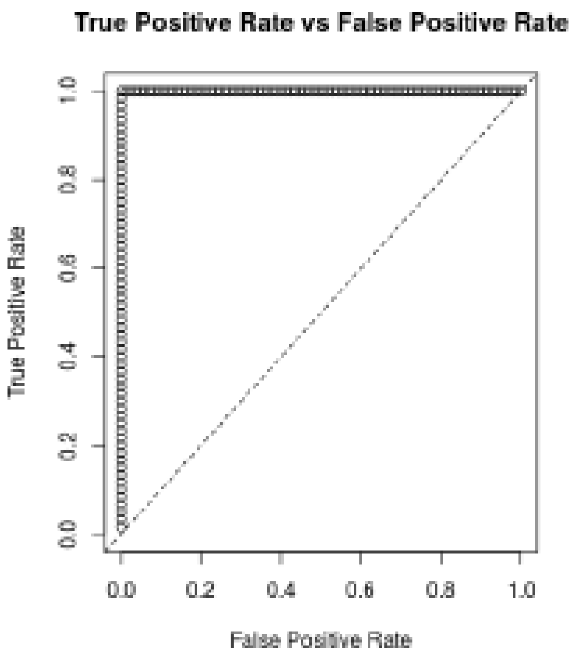
Autoencoder: Different number of neurons

Classifier: 4 layers with 10 neurons each

# Processed by only the largest autoencoder



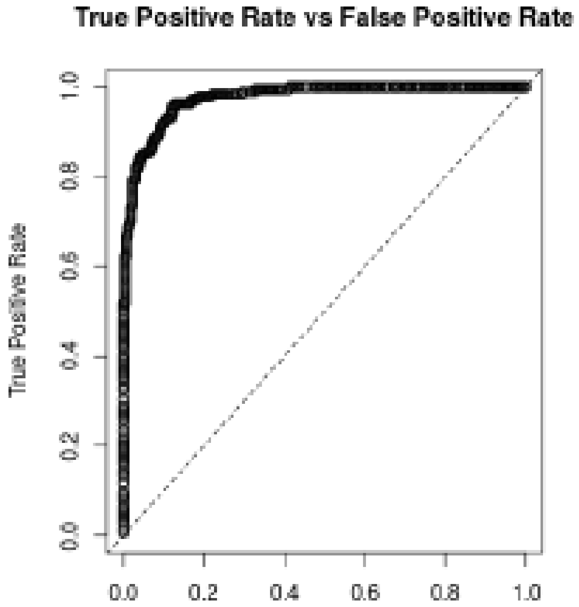
Processed by autoencoder
hidden unit =2000



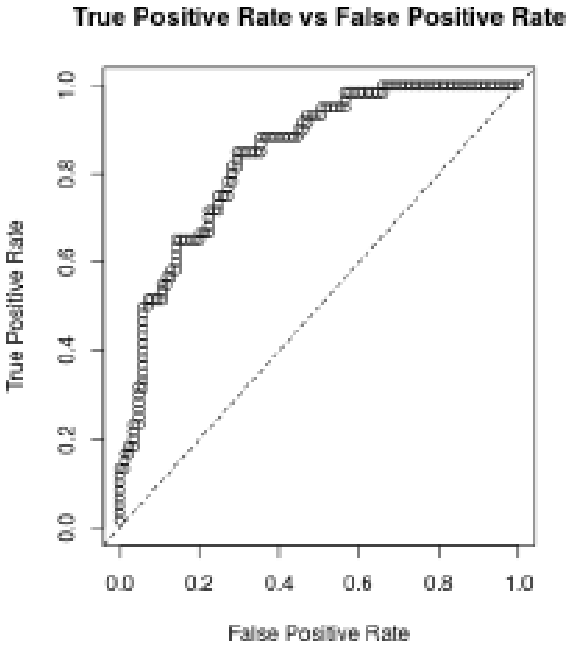
No processing
with 4666 SNPs

| AUC | 0.9969 | 0.9998 |
|---|---|---|

# Processed by all autoencoders



Processed by all autoencoders
smallest hidden unit =50



No processing
with 51 SNPs

| | | |
|---|---|---|
| AUC | 0.9745 | 0.8354 |

# Significance of this study

- Information contained in 4666 SNPs can be compressed (gradually) to be represented by output of as few as 50 neurons
- Such representation performed better than selecting SNPs by considering them separately based on certain p-value cut-off

# Comment

- This paper demonstrated neural network can be used for feature selection
    - ➢ But still suffers from the black-box character of this method
    - ➢ Authors claimed the advantage of stacked autoencoder is to “capture nonlinear dependencies and epigenetic interactions.”
        - ➢ How would it compare with other feature selection methods like principal component analysis?
- Classification task with neural network
    - ➢ The performance seemed surprisingly good with this sample size
    - ➢ Comparison with other methods in the same cohort is needed
- Reproducibility
    - Methods were not fully disclosed e.g. the hyper-parameters used in training the neural network
    - Authors did not explain clearly how was the classification done

- The End