Journal Club

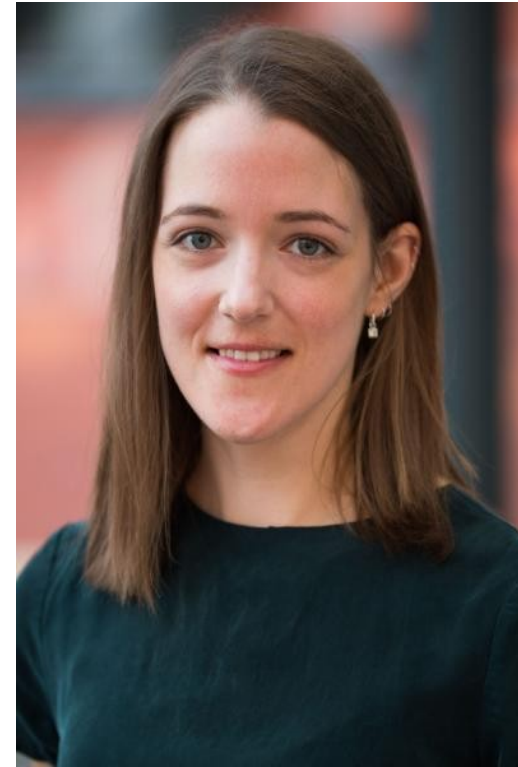https://github.com/richelbilderbeek/
journal_club_20220120

Why this article

Google 'GWAS + "machine learning"' for 2021

Start of collaboration

Carl Nettelblad

Kristiina Ausmees

Background

Can we order genomes?

Similar genomes close

Because then we can …

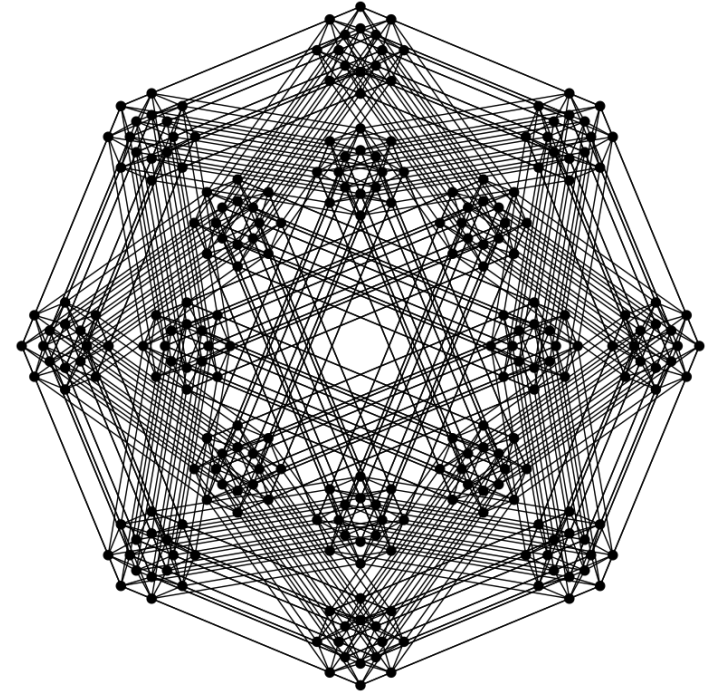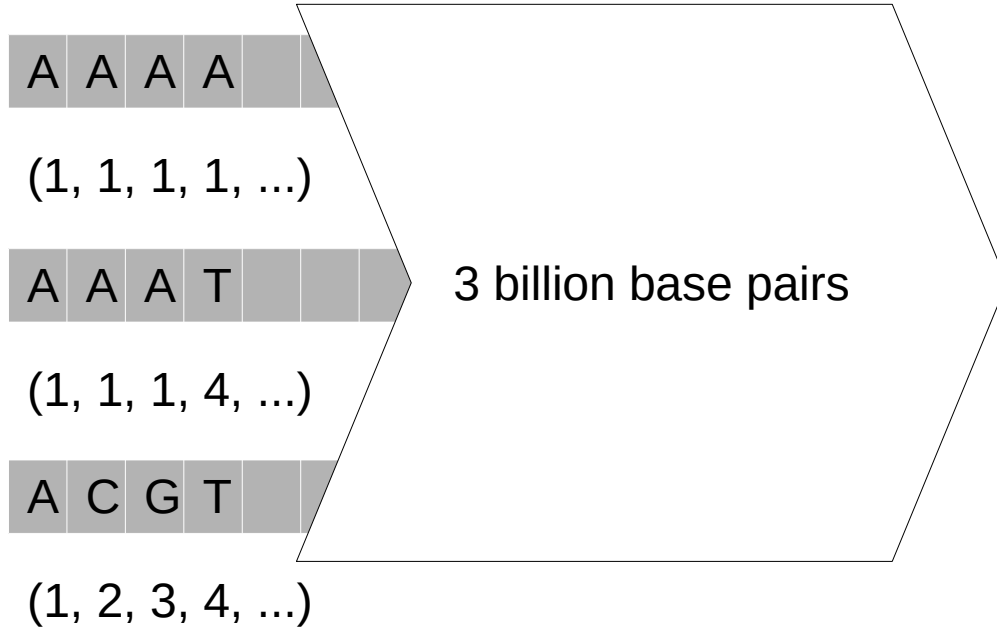See amount of genetic variation

Identify population structure

Do ancestry mapping

Simulate genotypes

It is easy to order genomes

A A A A

(1, 1, 1, 1, ...)

A A A T

(1, 1, 1, 4, ...)

A C G T

(1, 2, 3, 4, ...)

3 billion base pairs

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Background

Can we order genomes in a <u>smart</u> way?

Similar genomes close

Because then we can ...

See amount of genetic variation
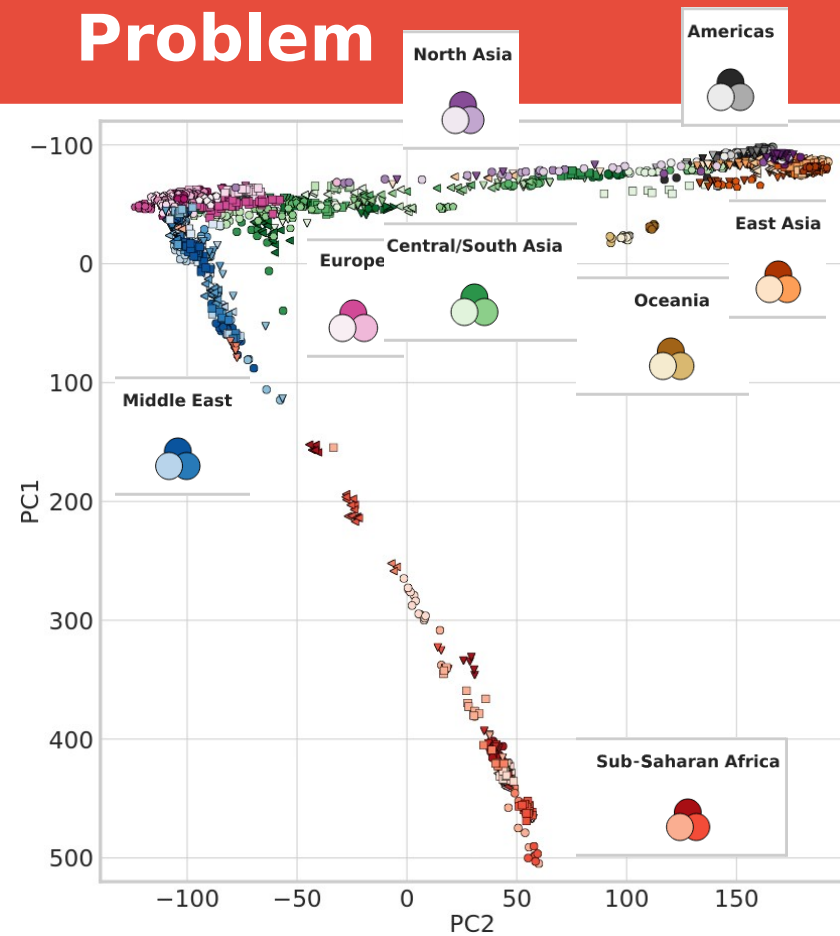
Identify population structure

Do ancestry mapping

Simulate genotypes

Dimensionality reduction

Problem

PCA reduces dimensionality linearily

Typical V shape

Overlap in first 2 dimensions

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Hypothesis

Convolutional autoencoders seem promising

Can work with noisy data as well
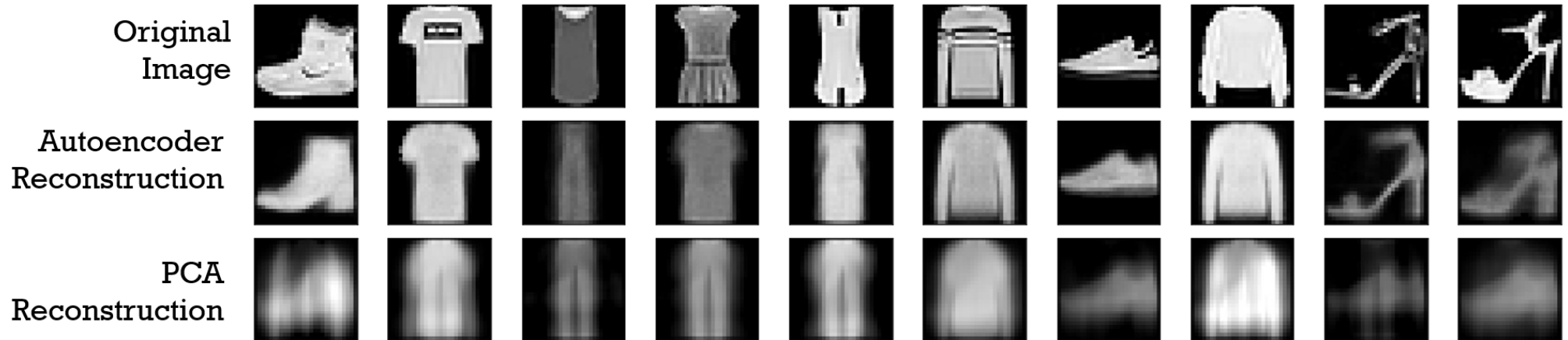
Set up a convolutional autoencoder

See how well it performs

NEW

75 75 2 75 75

| conv | conv | conv | maxpool | conv | | conv | upsample | conv | conv | conv | conv |
| 8 | 8 | 8 | 2 | 8 | | 8 | 2 | 9 | 9 | 8 | 1 |

fully-connected layers

encoder

decoder

Neural networks

Auto encoders

Convolutional

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).
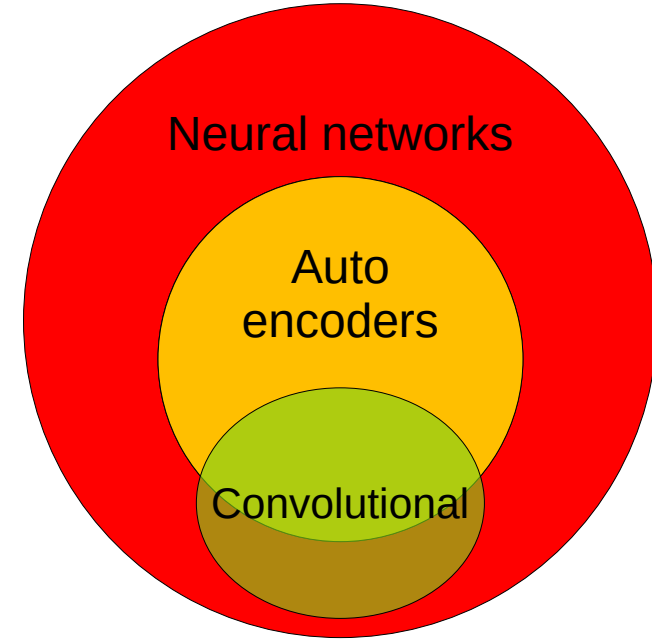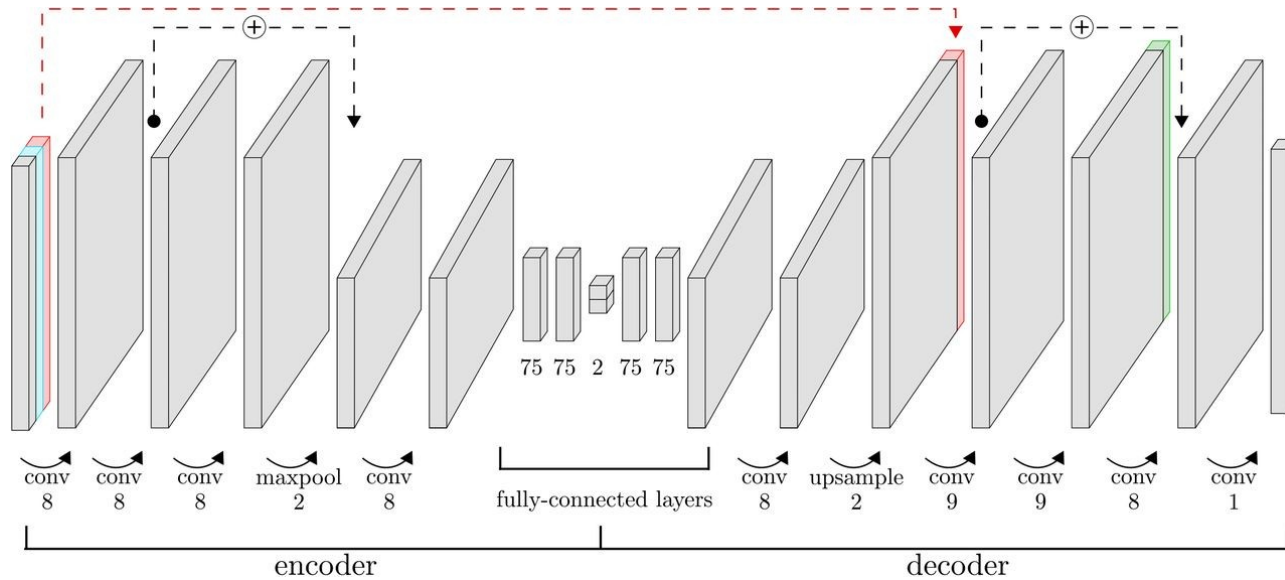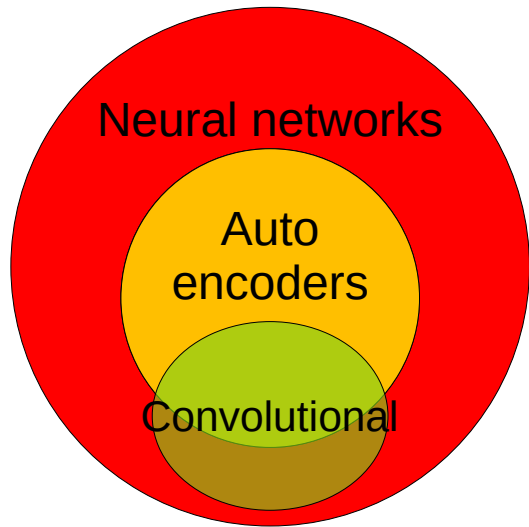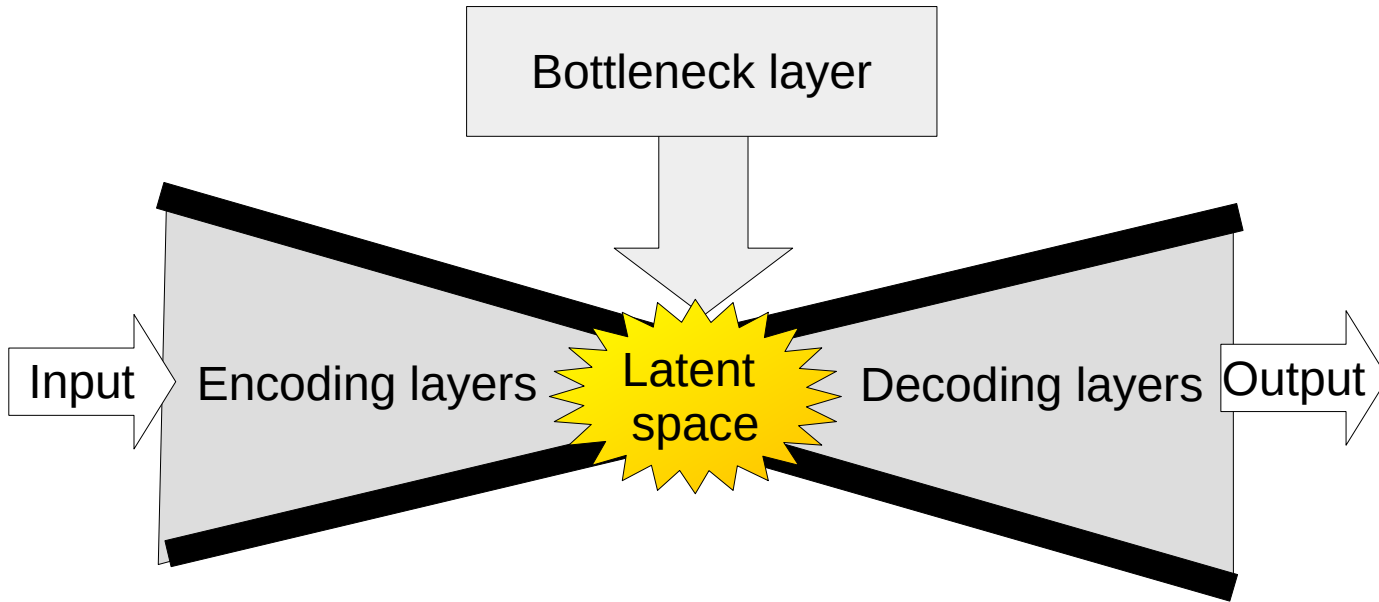
What is a convolutional autoencoder?

Convolutional: preprocess the data, i.e. do not only use the raw data

Autoencoder: type of neural network that learns how to encode data

What is an autoencoder?

Bottleneck layer

Input

Encoding layers

Latent space

Decoding layers

Output

Goal: *learn* to represent/label data, nonlinearly

function $\underline{G}$ function $\underline{H}$

y_1
y_2
$\vdots$
y_m

y'_1
y'_2
$\vdots$
y'_m

INPUT LAYER   MAPPING LAYER   BOTTLE-NECK LAYER   DE-MAPPING LAYER   OUTPUT LAYER
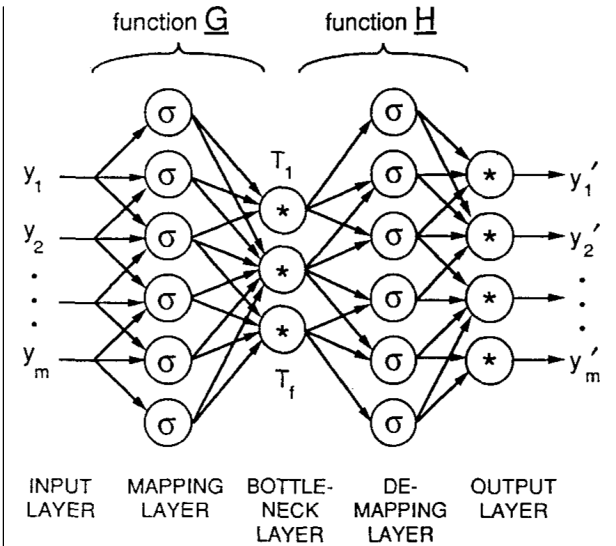
$Y = \{ y_1, y_2, \dots \}$
$Y' = \{ y'_1, y'_2, \dots \}$
Loss function = $E = Y - Y'$
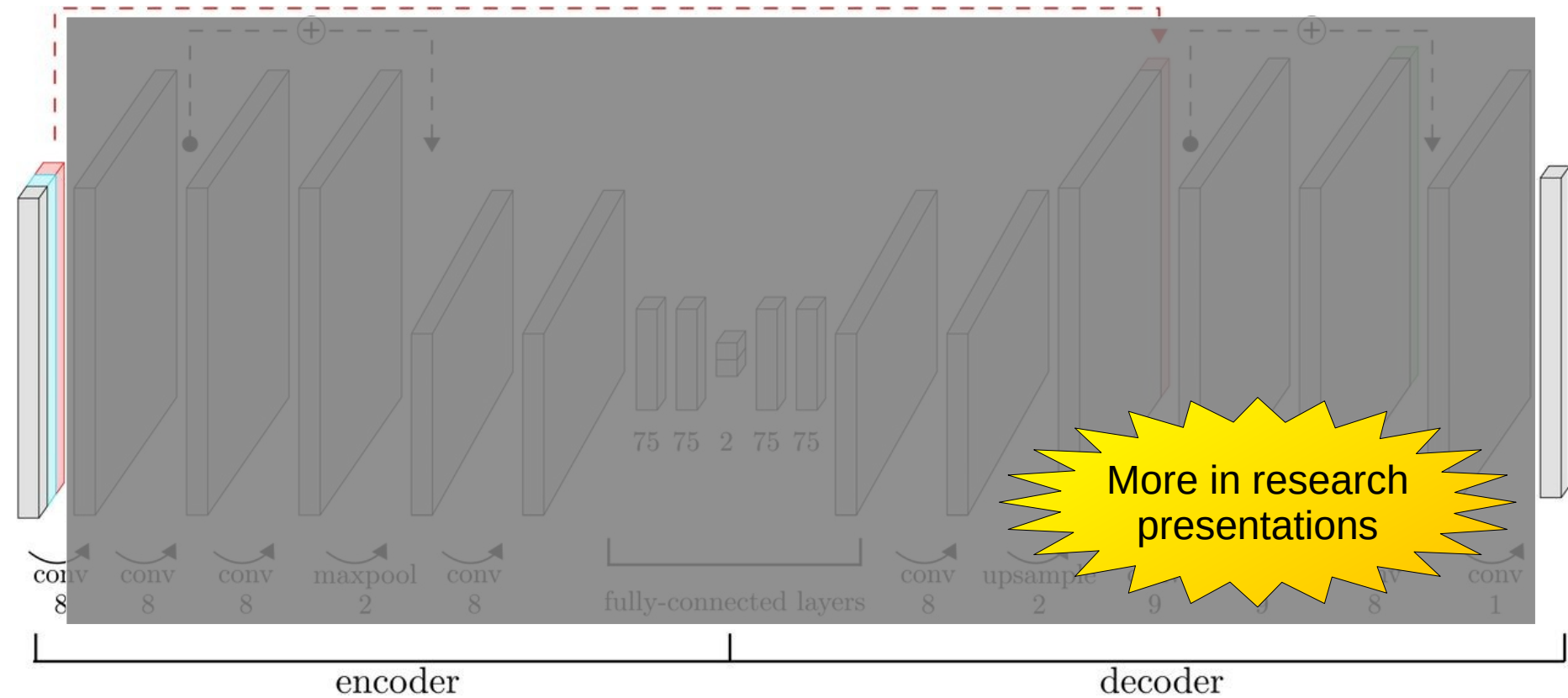
Kramer, Mark A. "Nonlinear principal component analysis using autoassociative neural networks." AIChE journal 37.2 (1991): 233-243.

GCAE

More in research presentations

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Article

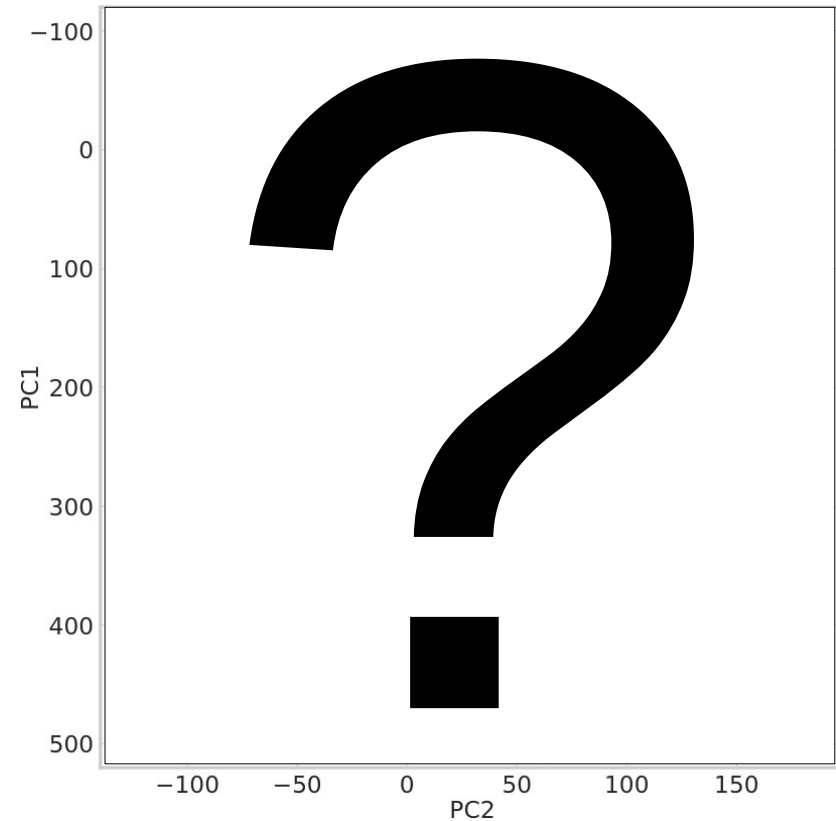## How well does GCAE perform?

genotype separation

deduce genotype origin

rare and common alleles

LD

Separating genotypes

Affymetrix Human Origins SNP array

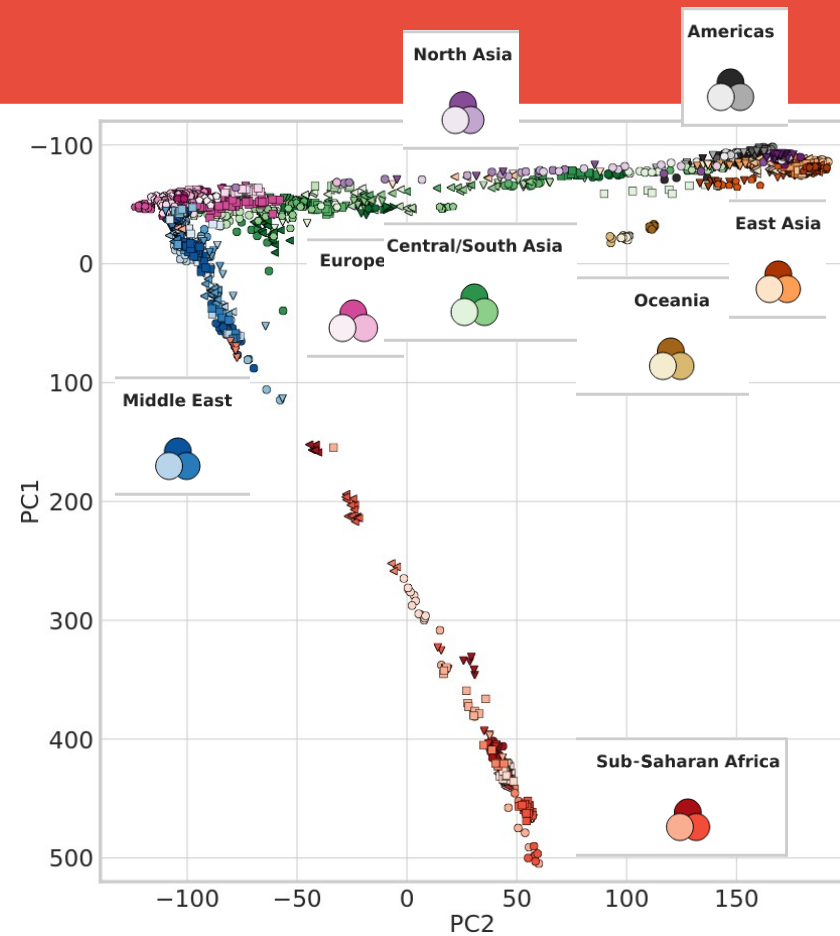**2,067 unrelated humans from 166 populations**

No sex chromosomes

Remove MAF < 1%

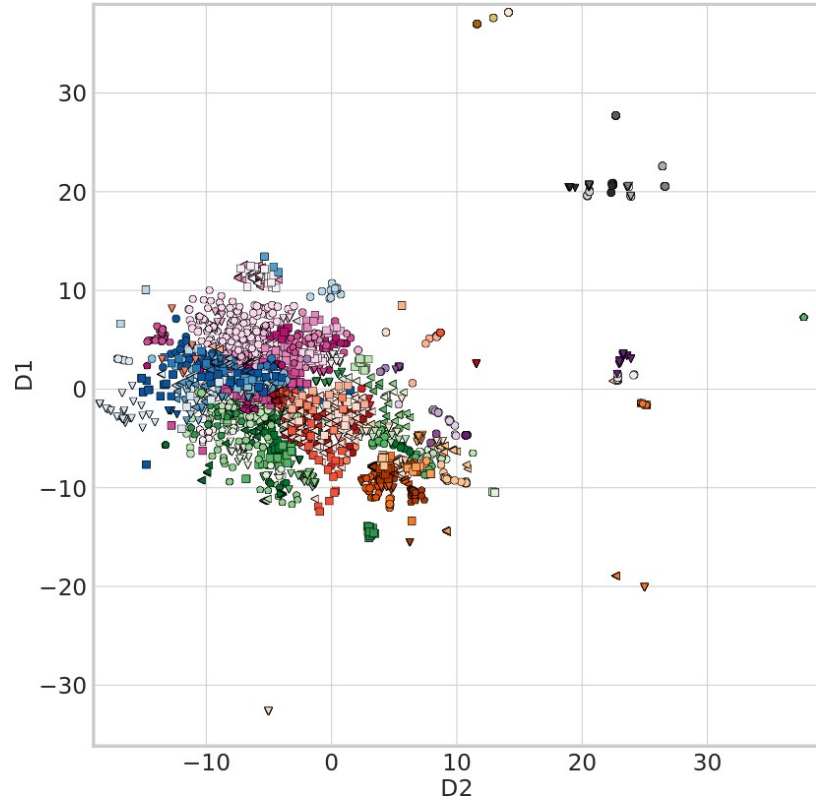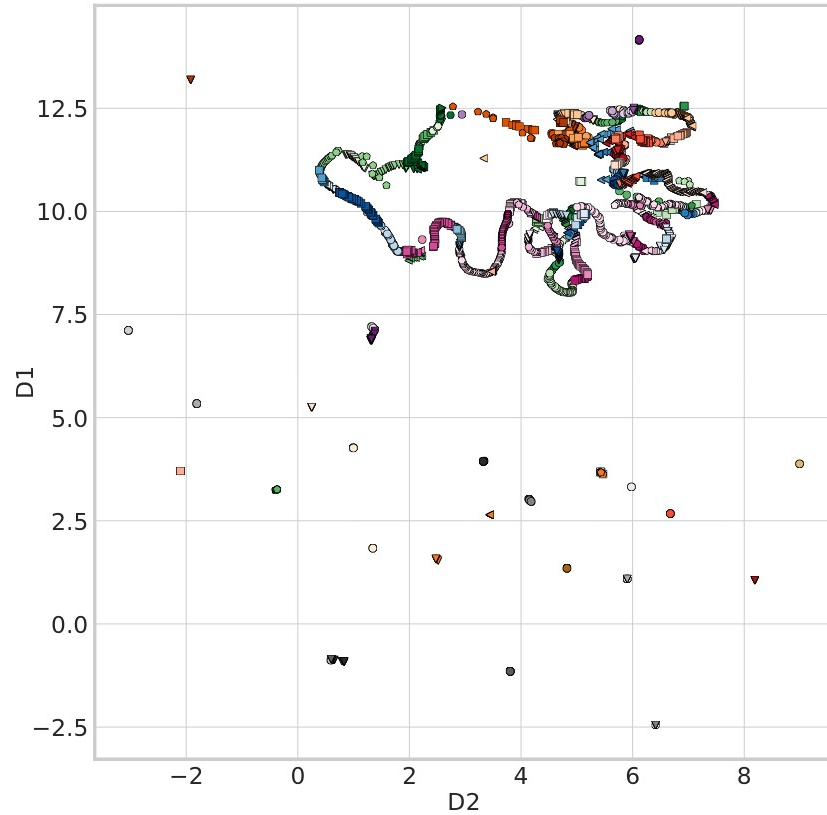**LD pruning, remove R^2 > 0.2**

160,858 biallelic sites

Show genotypes in first 2 PCs

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

PCA dimensionality reduction

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).
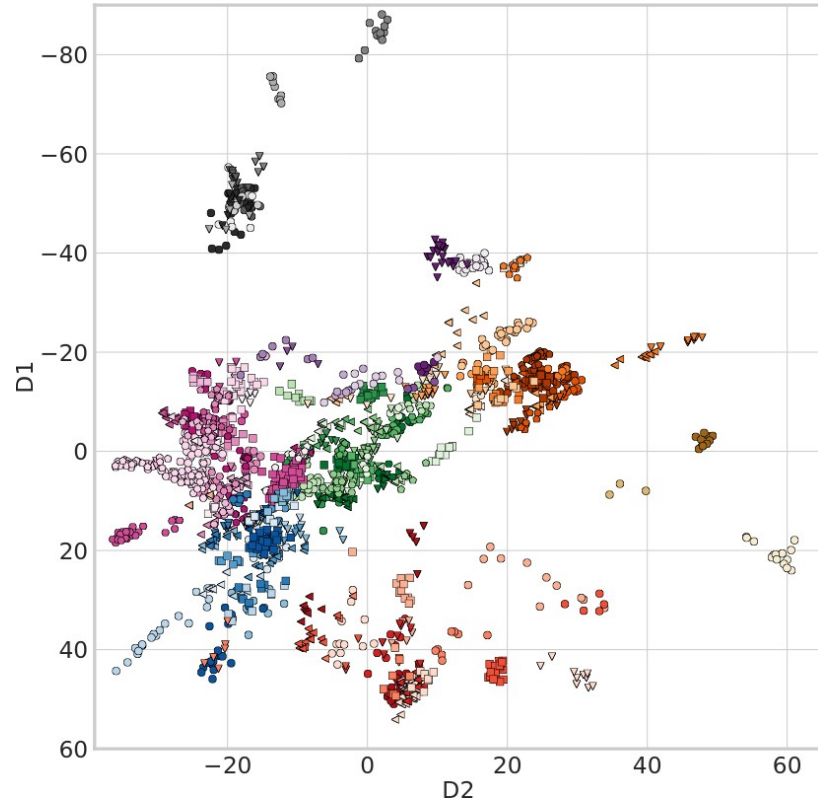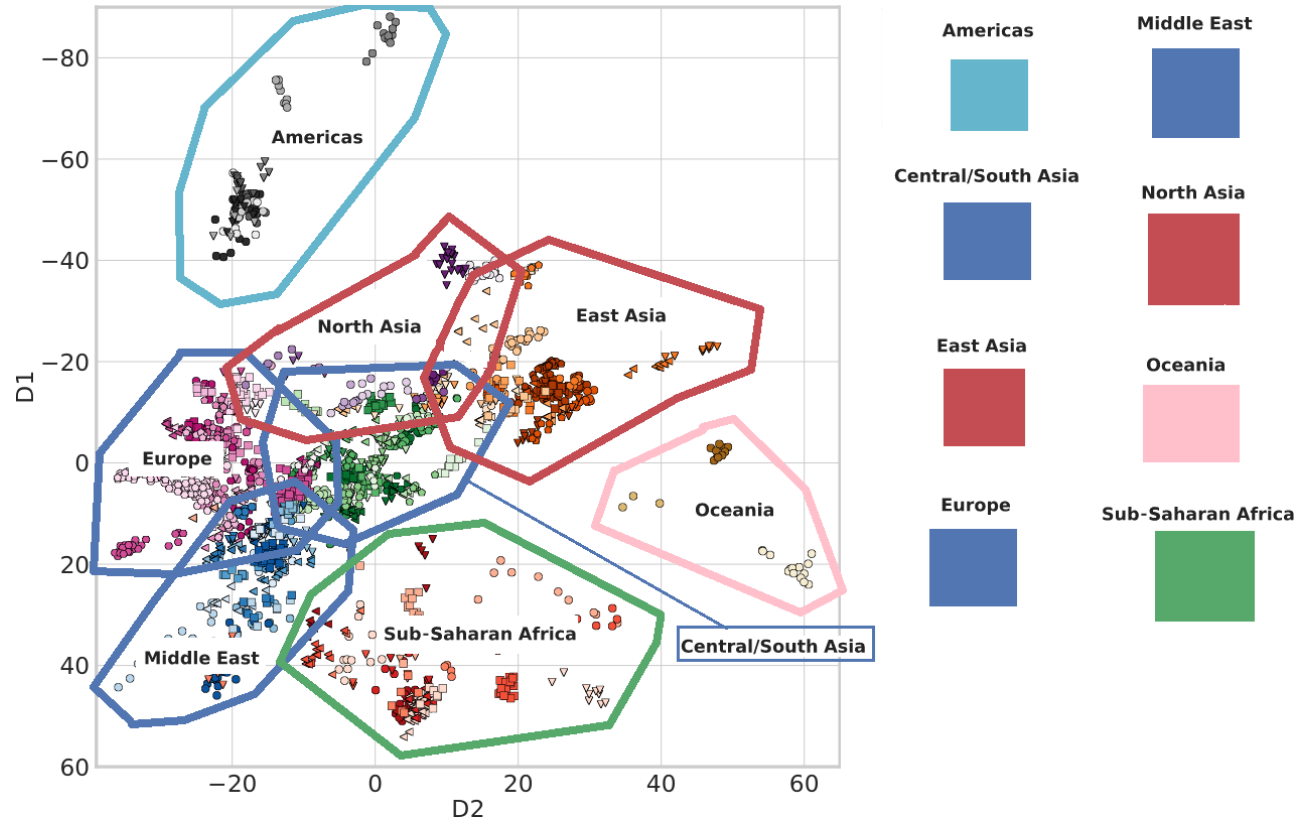
t-SNE dimensionality reduction

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

UMAP dimensionality reduction

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

GCAE dimensionality reduction

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

GCAE dimensionality reduction

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020). Adapted by Richel Bilderbeek

Questions

How well can genotypes be separated by eye?

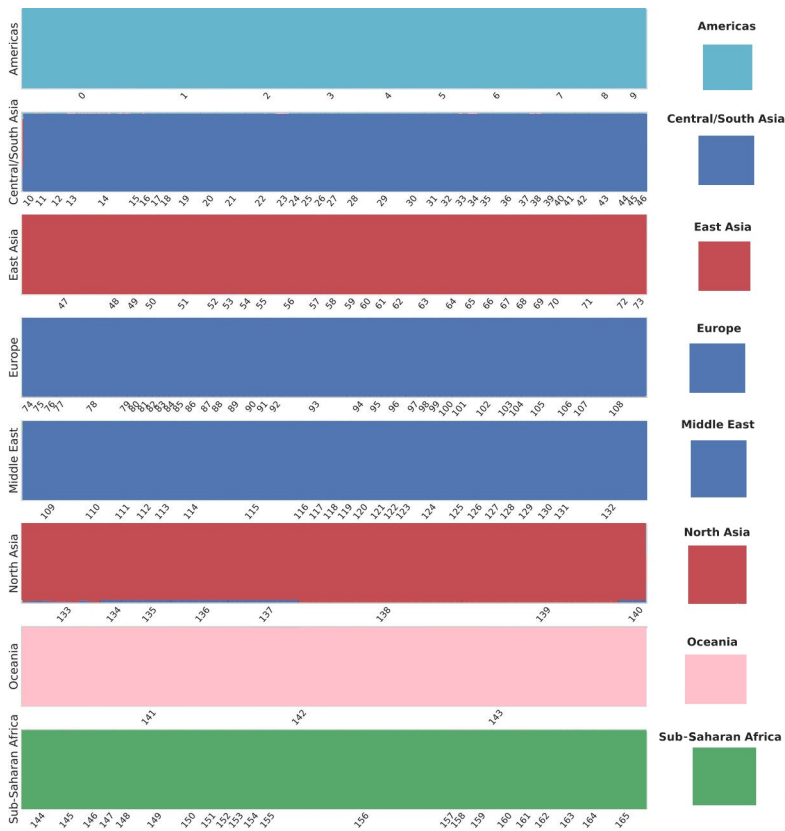## How well can we deduce the origin of a genotype?

Clustering similar to ADMIXTURE

F1 score

Other measurements

How well can we deduce the origin of a genotype?
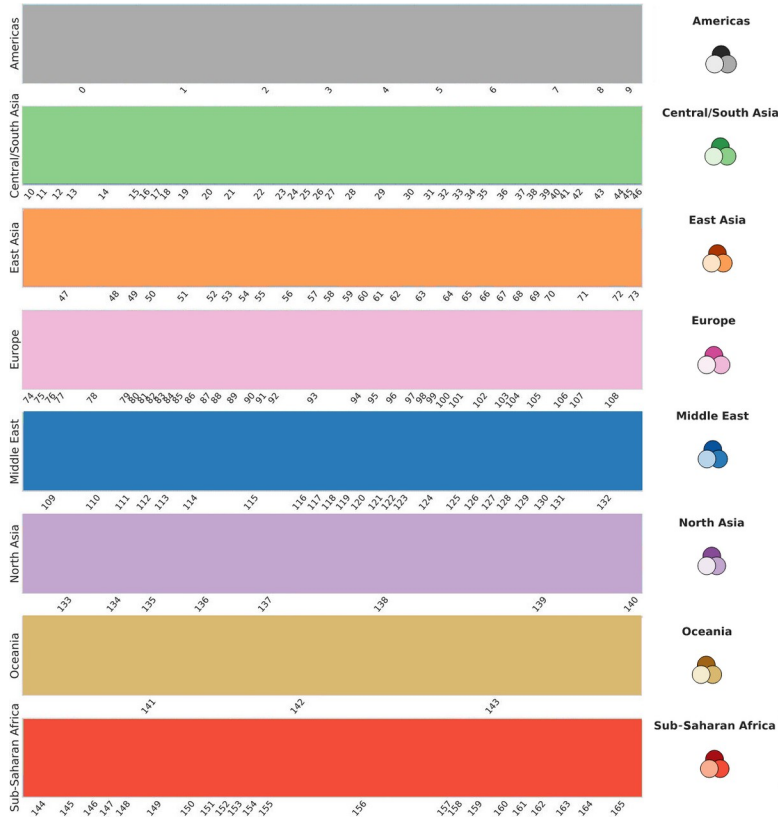
Go through all individuals (x axis)

Assign the likelihood it belong to a subcontinent (the rows)

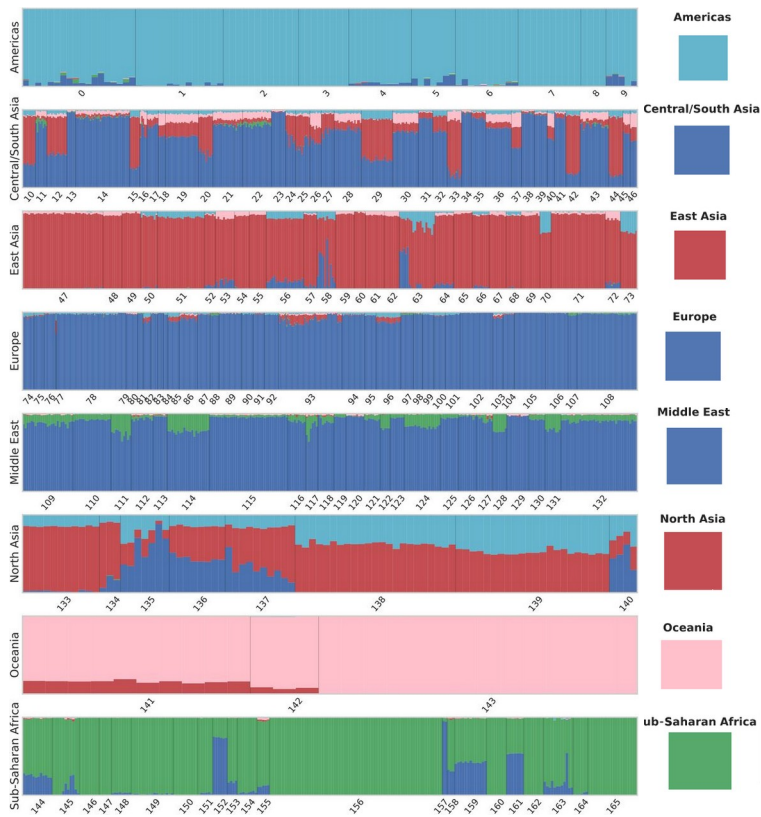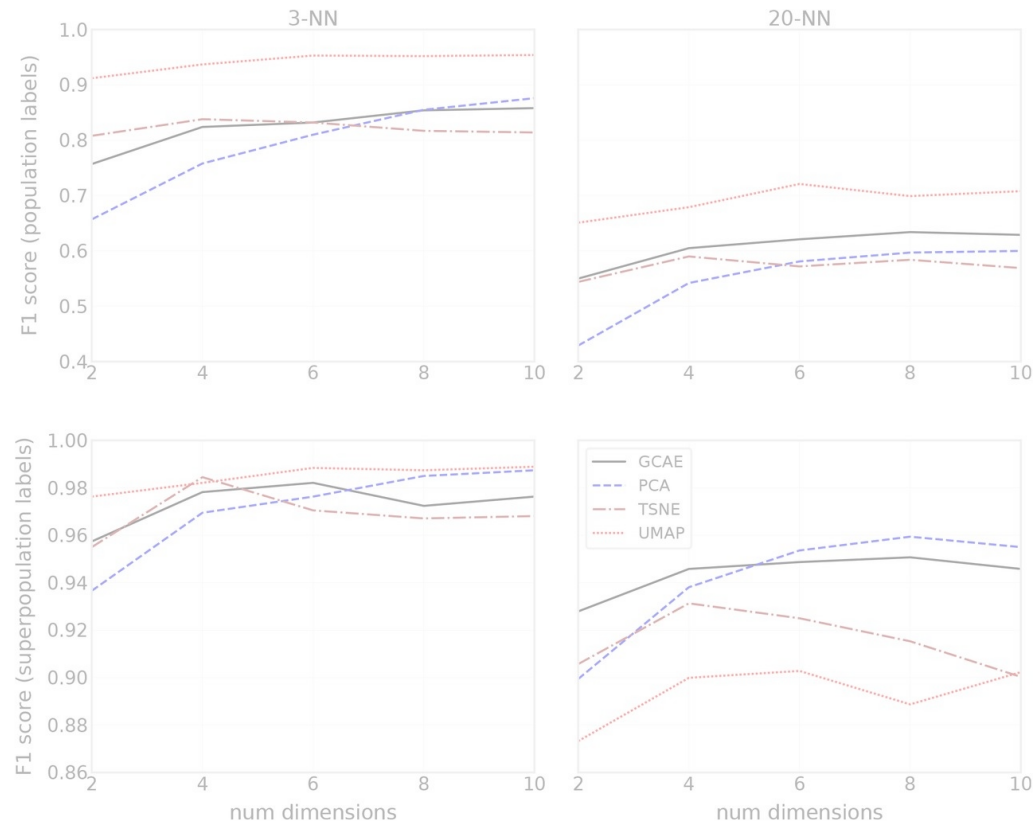Ideally, all individuals are assigned their correct subcontinent

How well can we deduce the origin of a genotype?

Go through all individuals (x axis)

Assign the likelihood it belong to a subcontinent (the rows)

Ideally, all individuals are assigned their correct subcontinent

Use five clusters
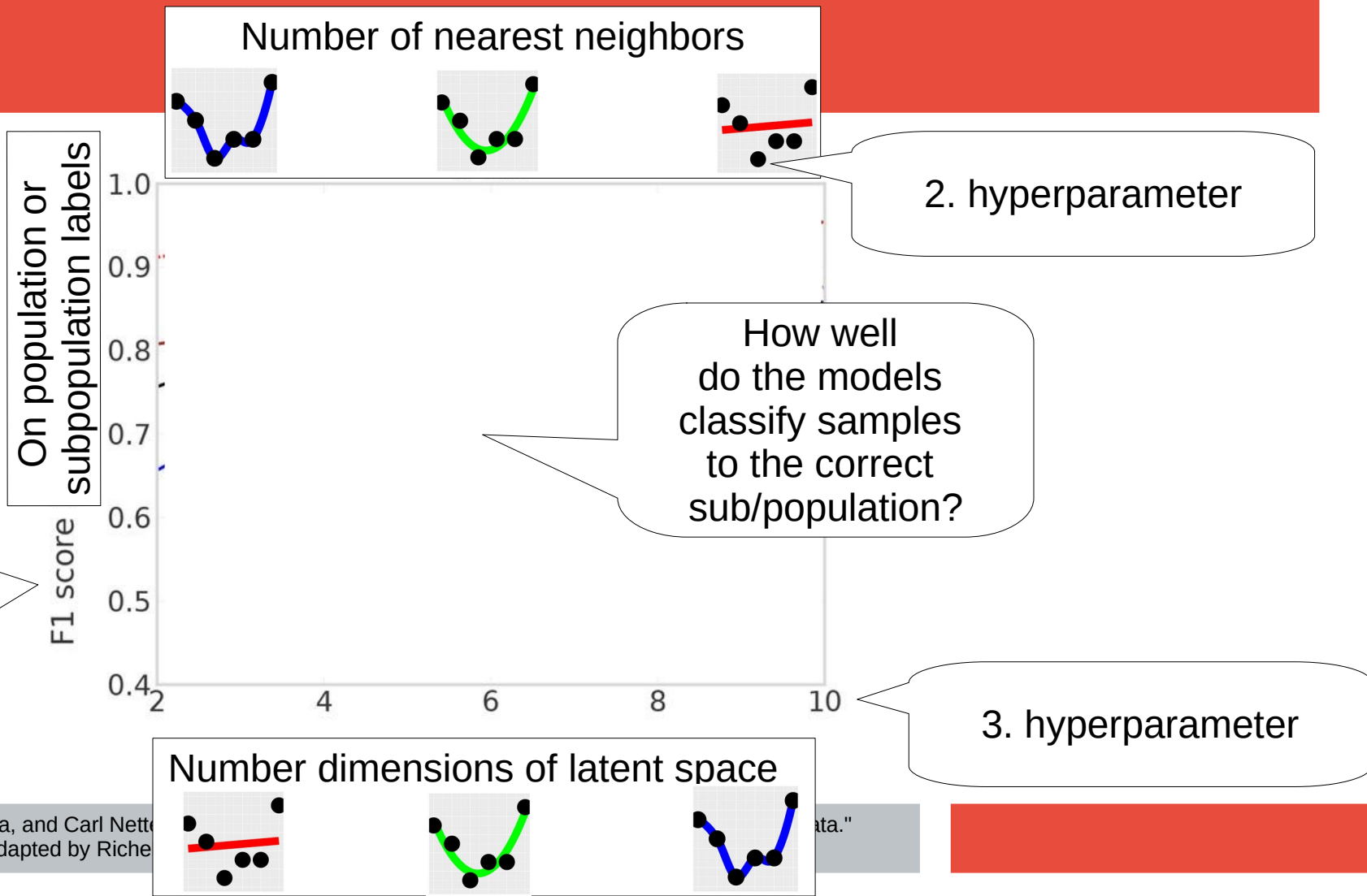
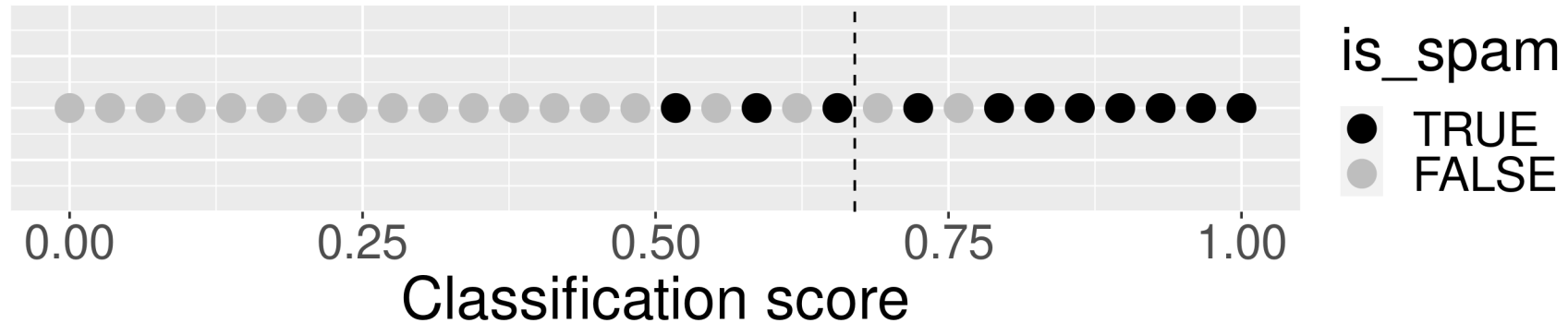# GCAE: ADMIXTURE ↔ plot
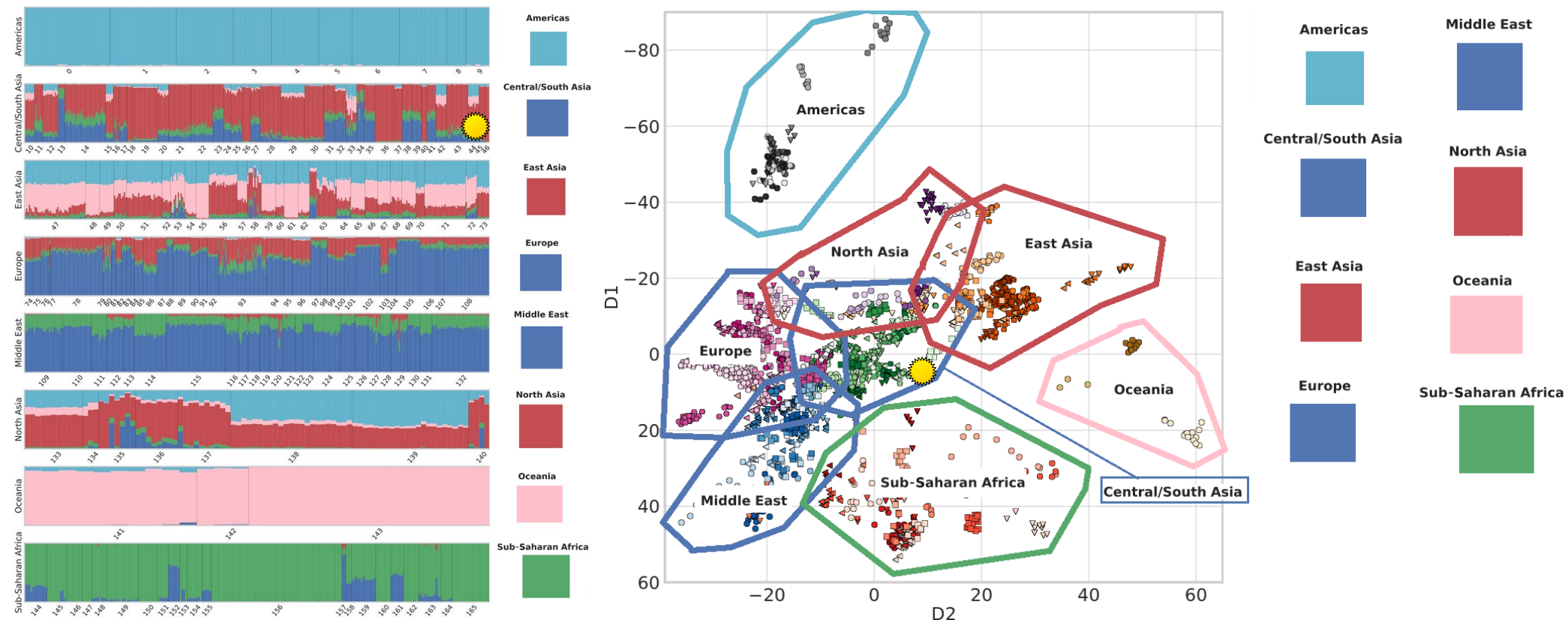
PCA: ADMIXTURE
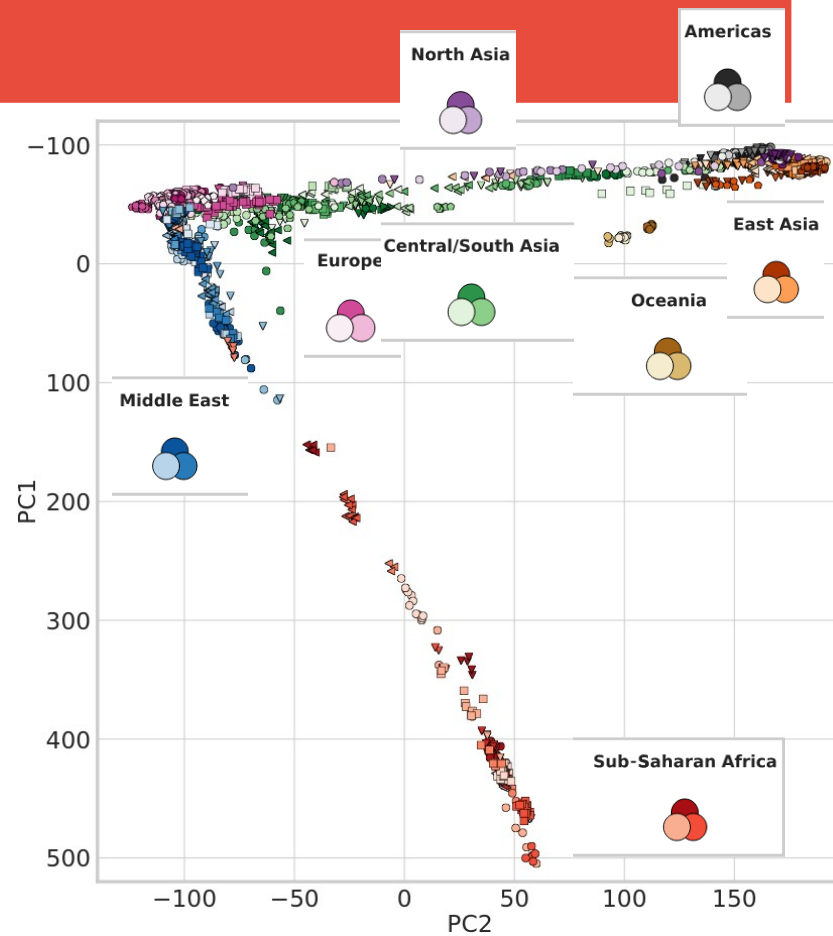
Figure 4

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Figure 4

What is the F1 score?

A metric for the trade-off between precision and recall →

What is precision?

Precision:

how often is the classifier right?

the number of true positives of all *estimated* positives
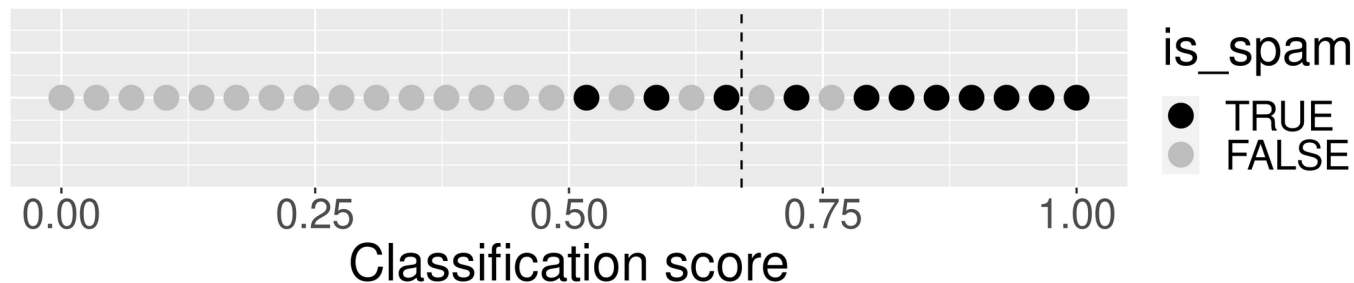
precision

$= n_{tp} / (n_{tp} + n_{fp})$

$= 8 / (8 + 2) = 8/10$

where

n_{tp} = # of true positives

n_{fp} = # of false positives

is_spam

● TRUE

● FALSE

Classification score

What is recall?

Recall:

how often are the positive cases recognized as such?
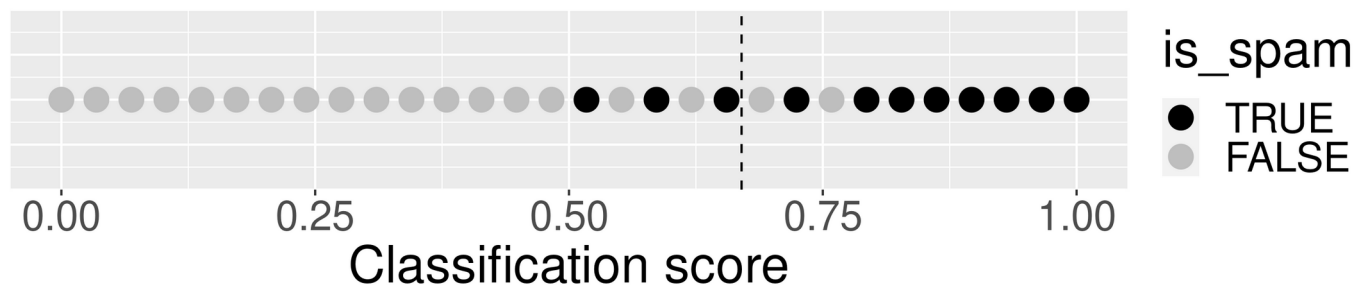
the number of true positives of all *known* positives

recall

$= n_{tp} / (n_{tp} + n_{fn})$
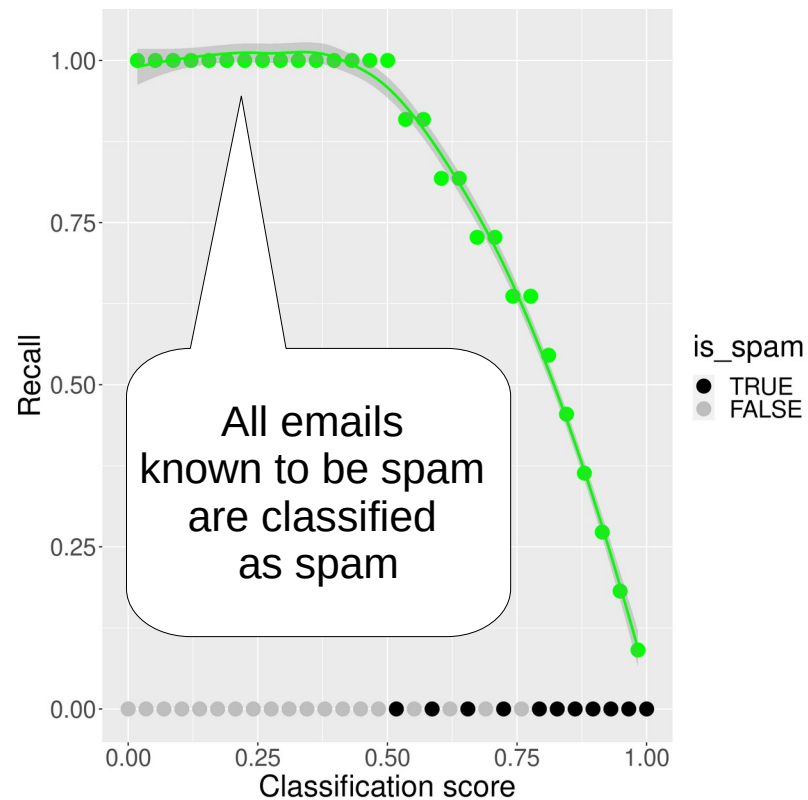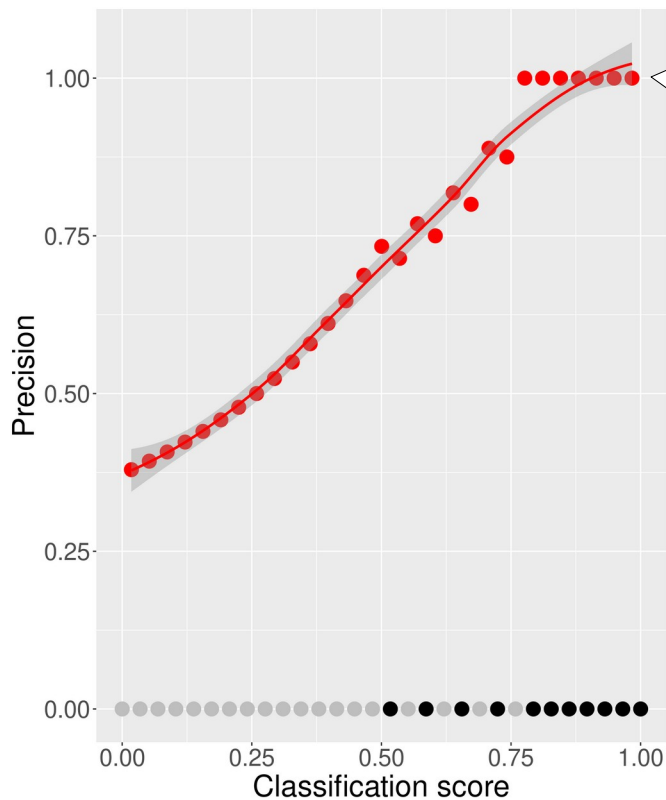
$= 8 / (8 + 3) = 8/11$

where

n_{tp} = # of true positives

n_{fn} = # of false negatives

Trade-off between precision and recall

https://github.com/richelbilderbeek/journal_club_20211216/f1_score.R

F1 Score

A metric for the trade-off between precision and recall:

The harmonic mean of precision and recall

F_1

$$= 2 / (precision^{-1} + recall^{-1})$$

$$= n_{tp} / (n_{tp} + \frac{1}{2}(n_{fp} + n_{fn}))$$

where

n_{tp} = # of true positives

n_{fp} = # of false positives
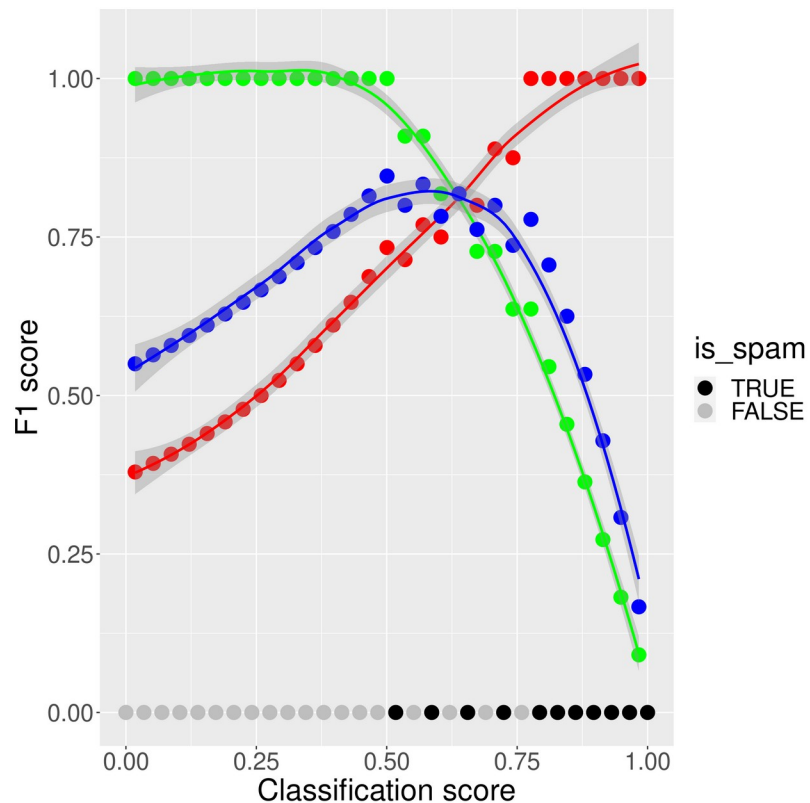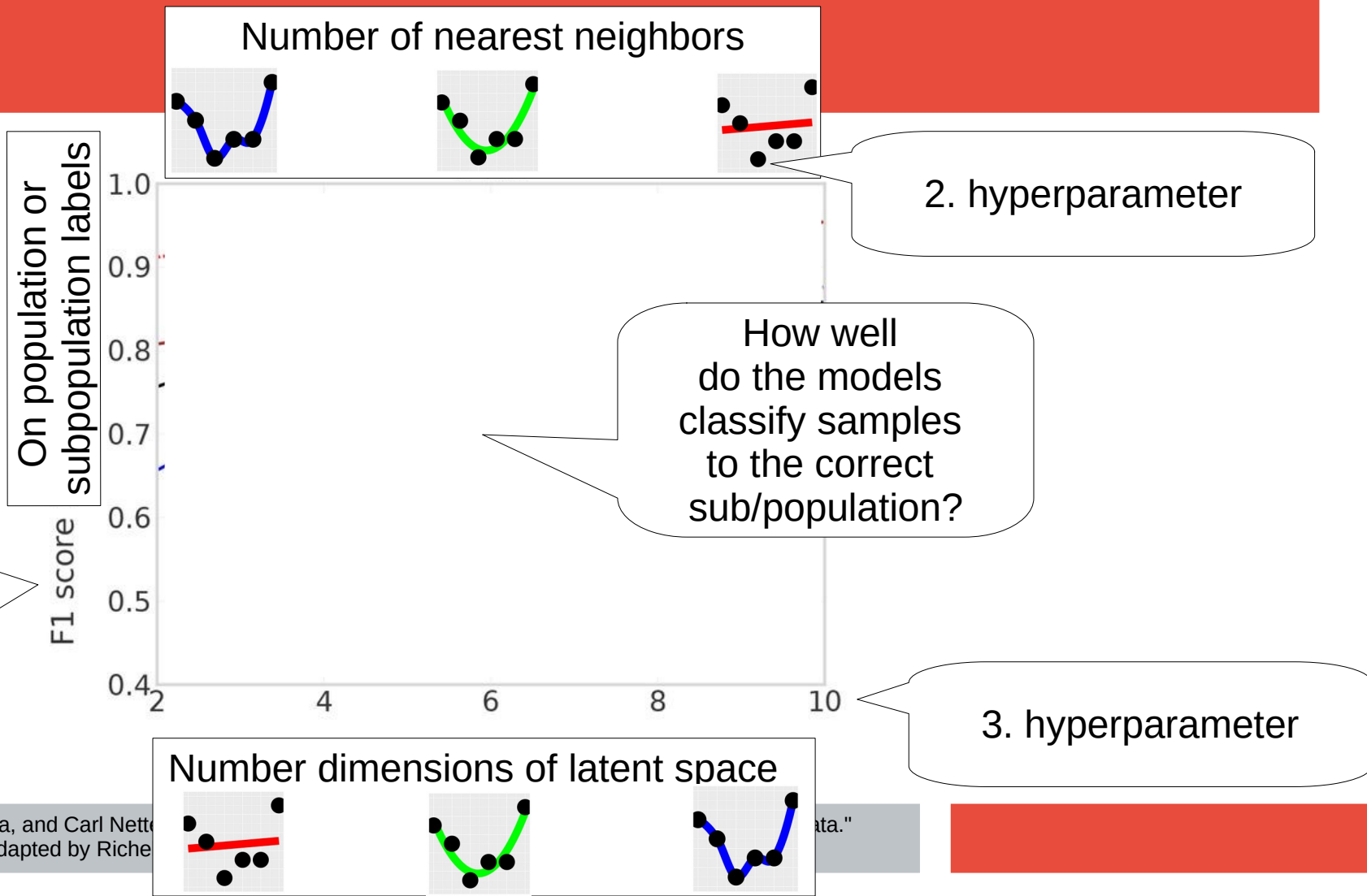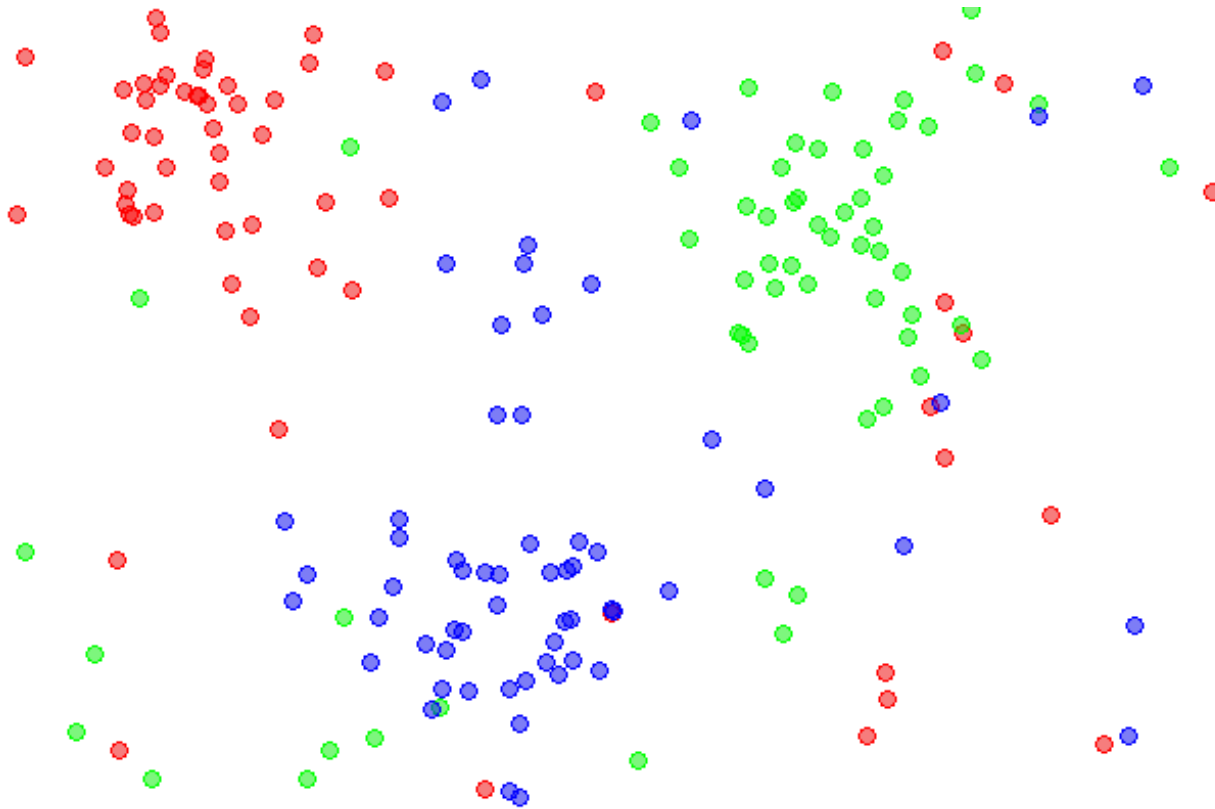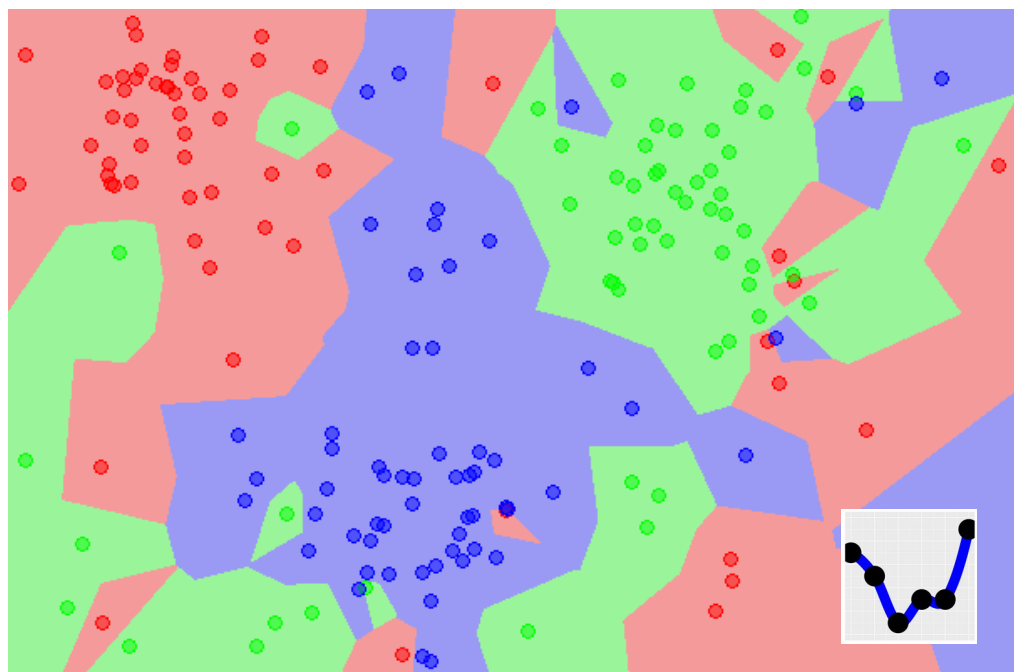
n_{fn} = # of false negatives

https://github.com/richelbilderbeek/journal_club_20211216/f1_score.R

Figure 4

Number of nearest neighbors

On population or subpopulation labels

F1 score

Number dimensions of latent space

1. performance score, higher is better

2. hyperparameter

How well do the models classify samples to the correct sub/population?

3. hyperparameter

Ausmees, Kristiina, and Carl Nette... ...ata." bioRxiv (2020). Adapted by Riche...

Number of nearest neighbors

https://en.wikipedia.org/wiki/File:Data3classes.png

Number of neighbors

1 5
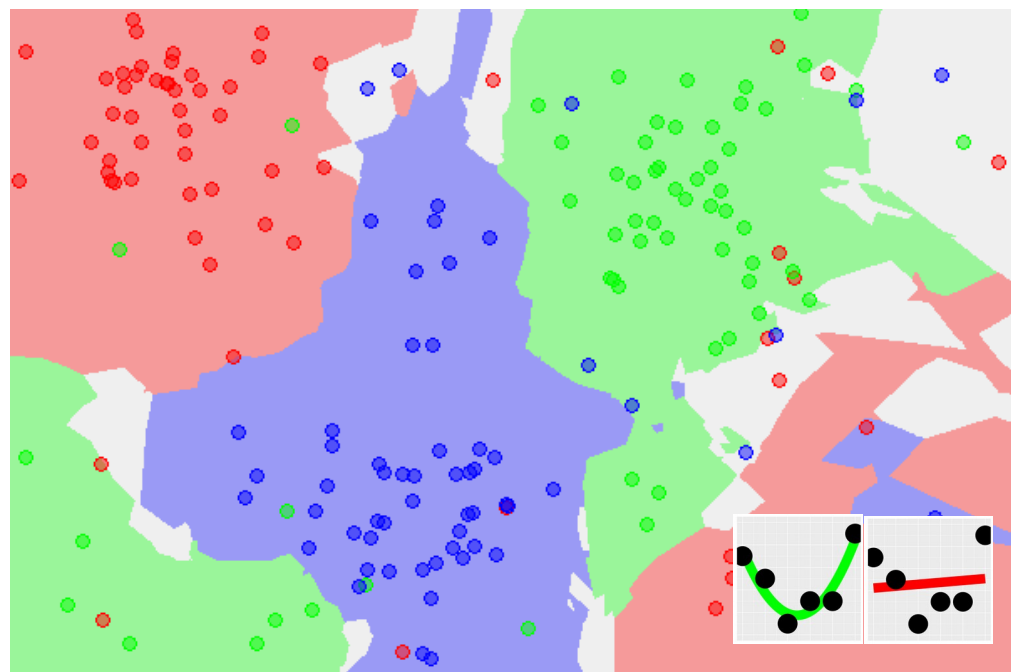
Optimal number for k?

Nontrivial to determine a priori

Prefer k being odd   √n   (√n)/2

stack**overflow**

Possible to determine from data
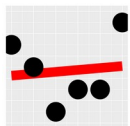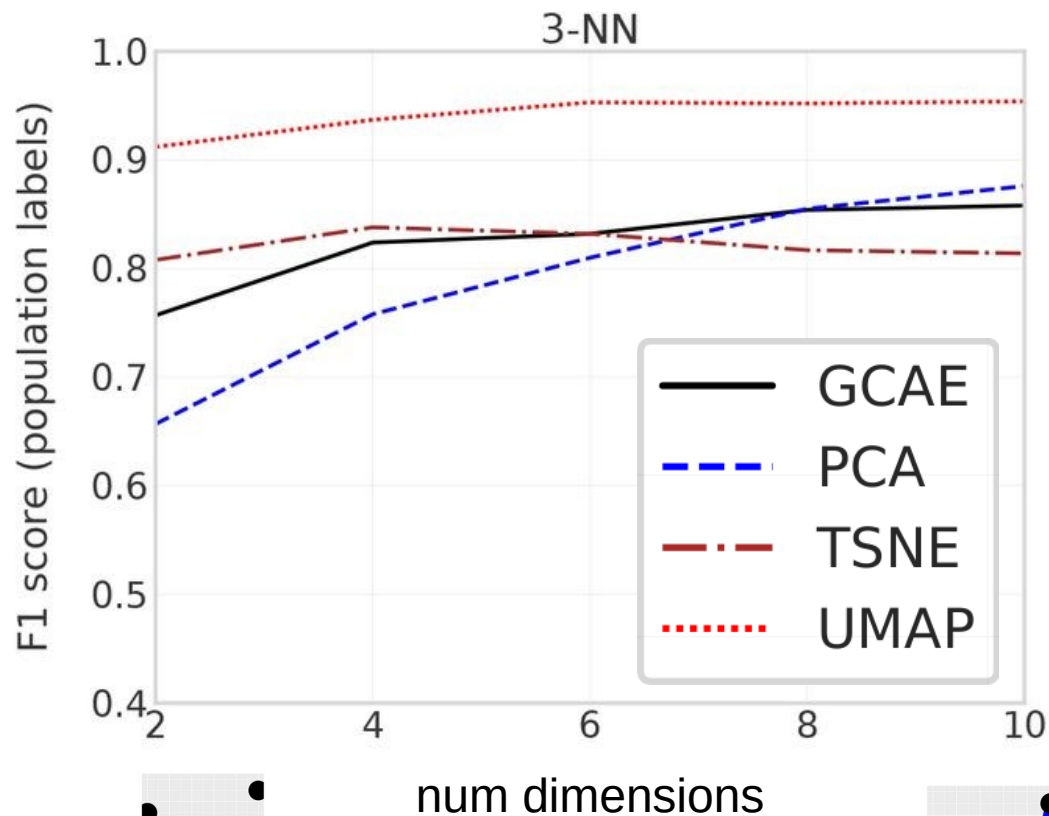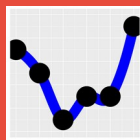
e.g. calculate the k with the lowest error

problem: these may differ between different methods (i.e. PCA, GCAE)

166 populations → √n ≈ 13, (√n)/2 ≈ **6**

8 superpopulations → √n ≈ 3, (√n)/2 ≈ **1**

k = number of neighbors
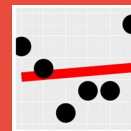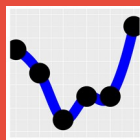n = the number of categories

Figure 4ab
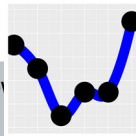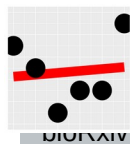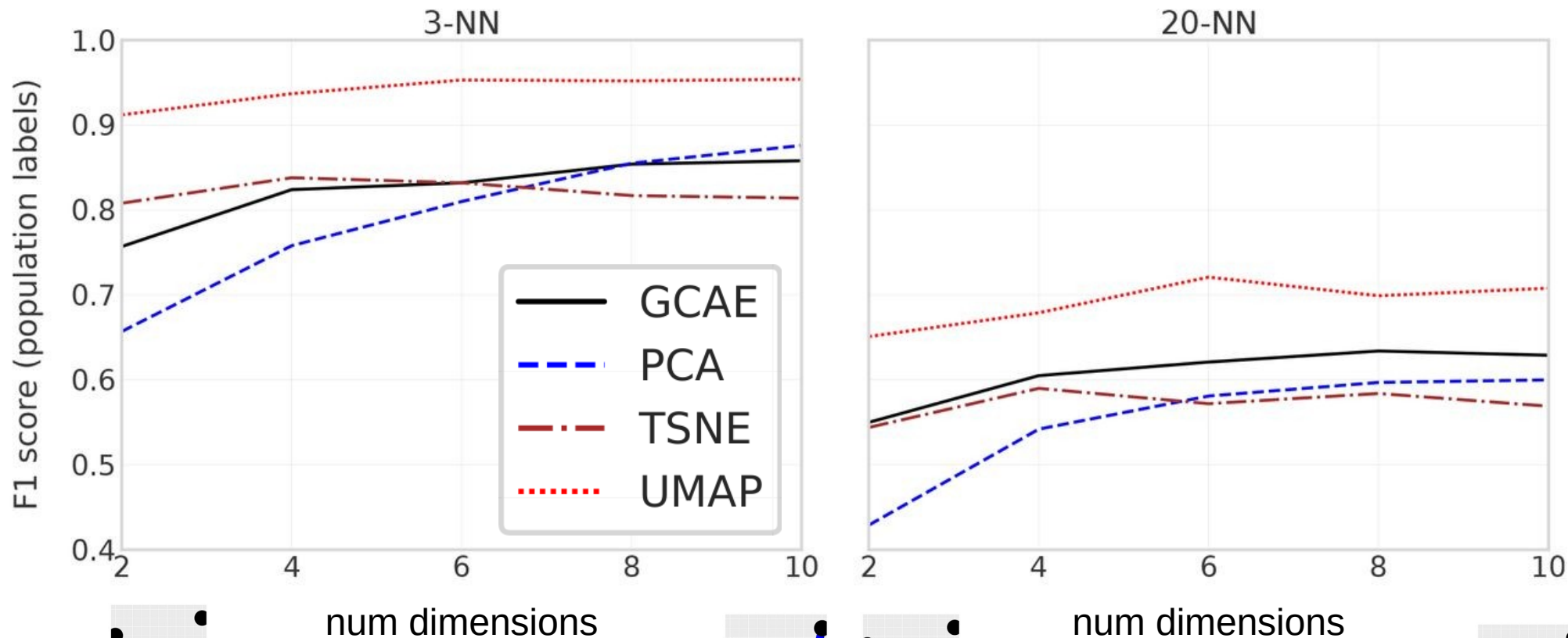
Figure 4 revisited

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).
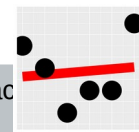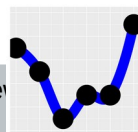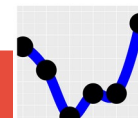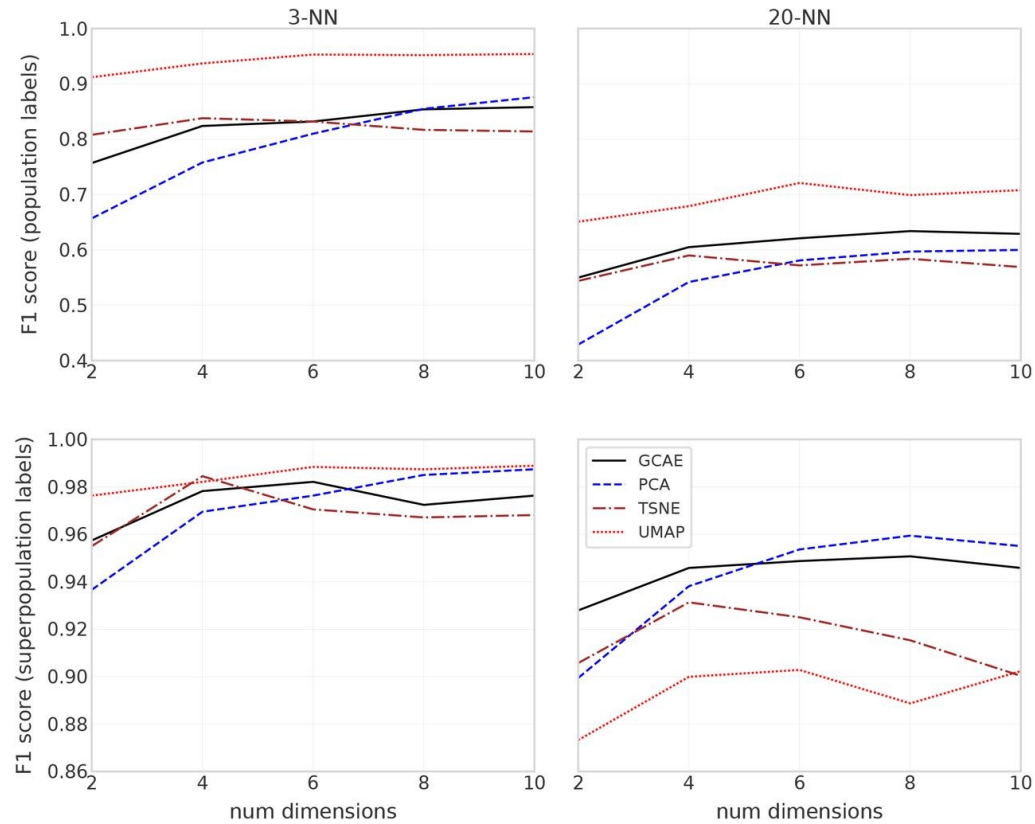
Questions

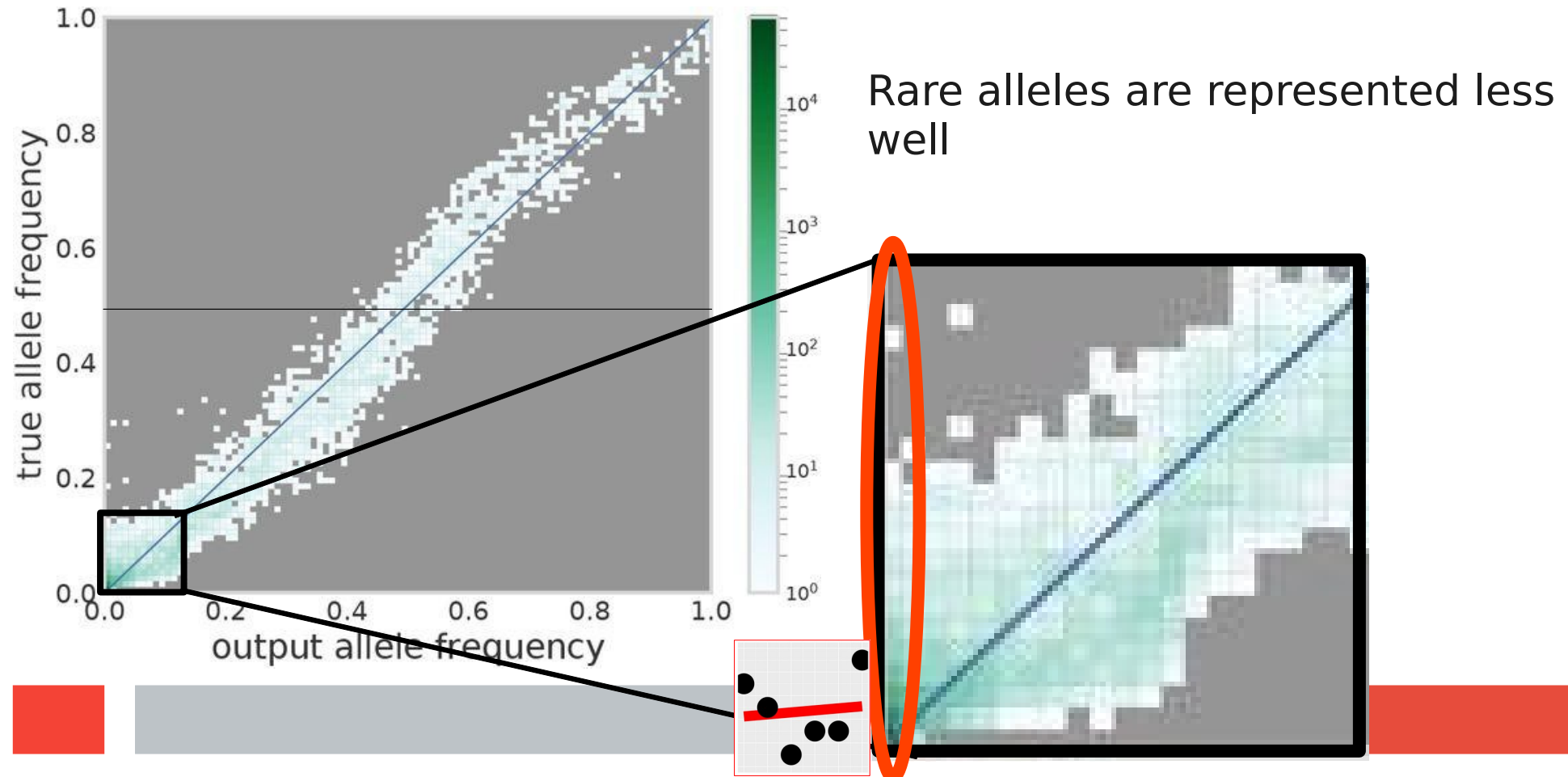How well can genotypes be separated by eye?

How well can we deduce the origin of a genotype?

Other measurements

Representing rare alleles

Representing linkage disequilibrium

Representing of rare alleles

Rare alleles are represented less well

Representing LD
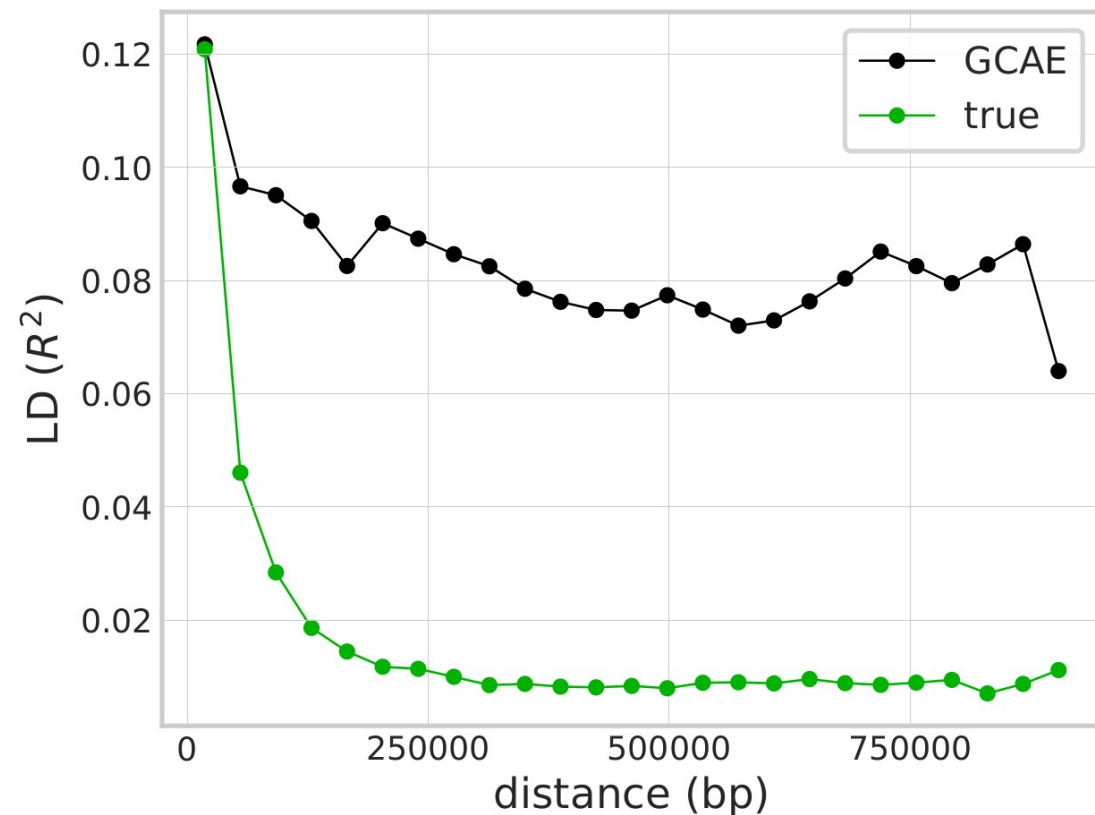
GCAE 'feels' all alleles are in LD

Proof it takes local structure into account

Pruned on LD < 0.2 ?

Allele independent

Discussion

Dimensionality reduction

ADMIXTURE

F1 scores

GCAE
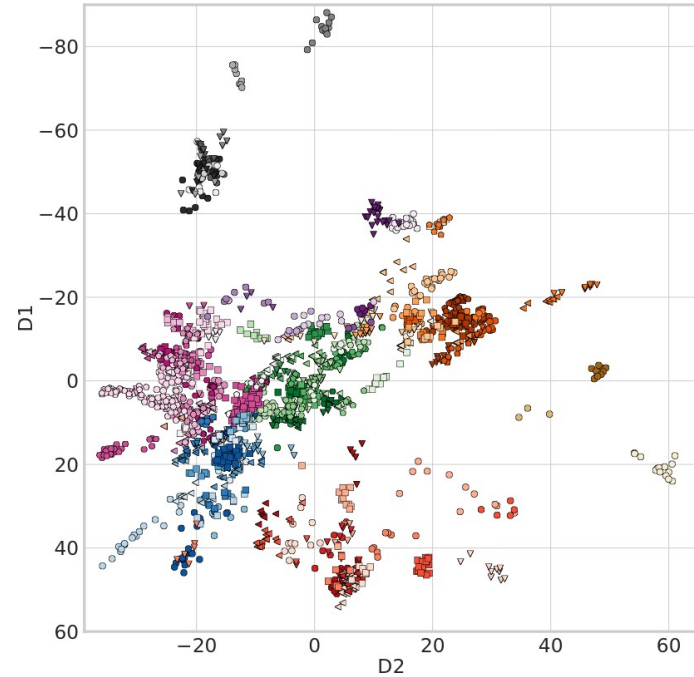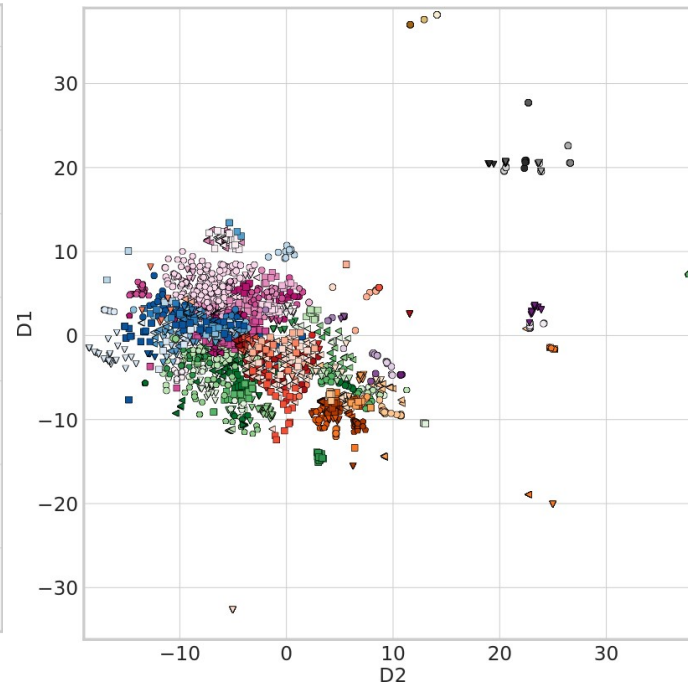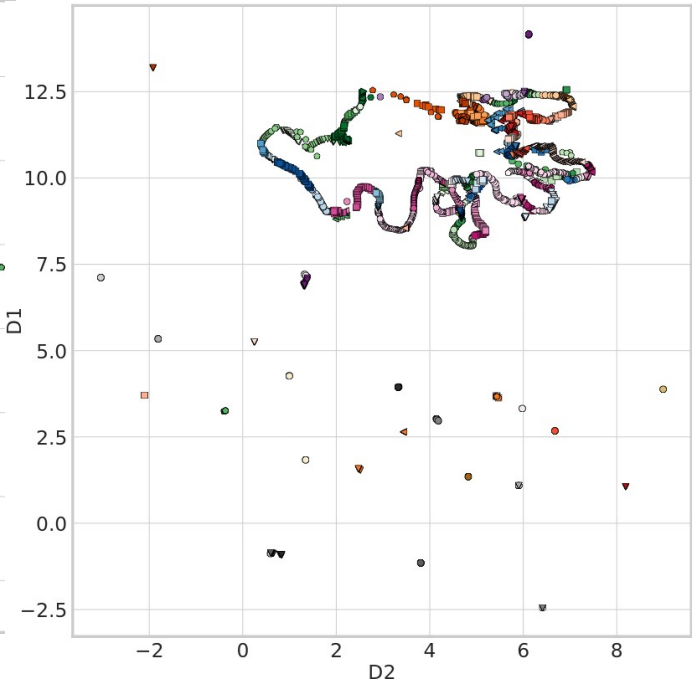
t-SNE

UMAP

GCAE versus UMAP

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Comparisons by other study
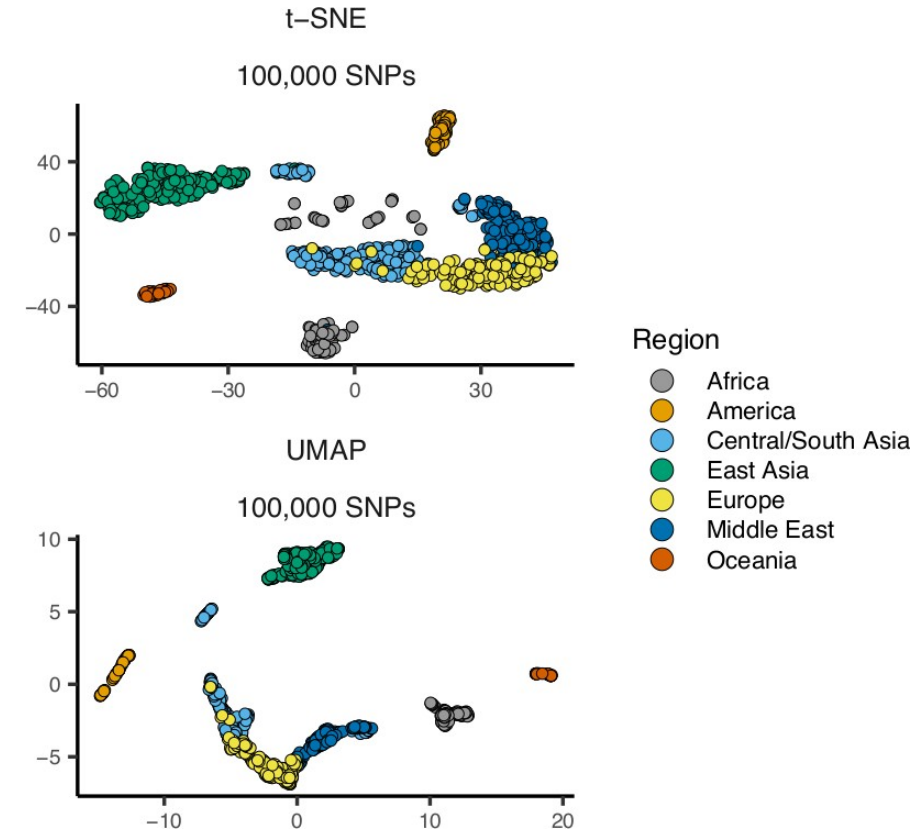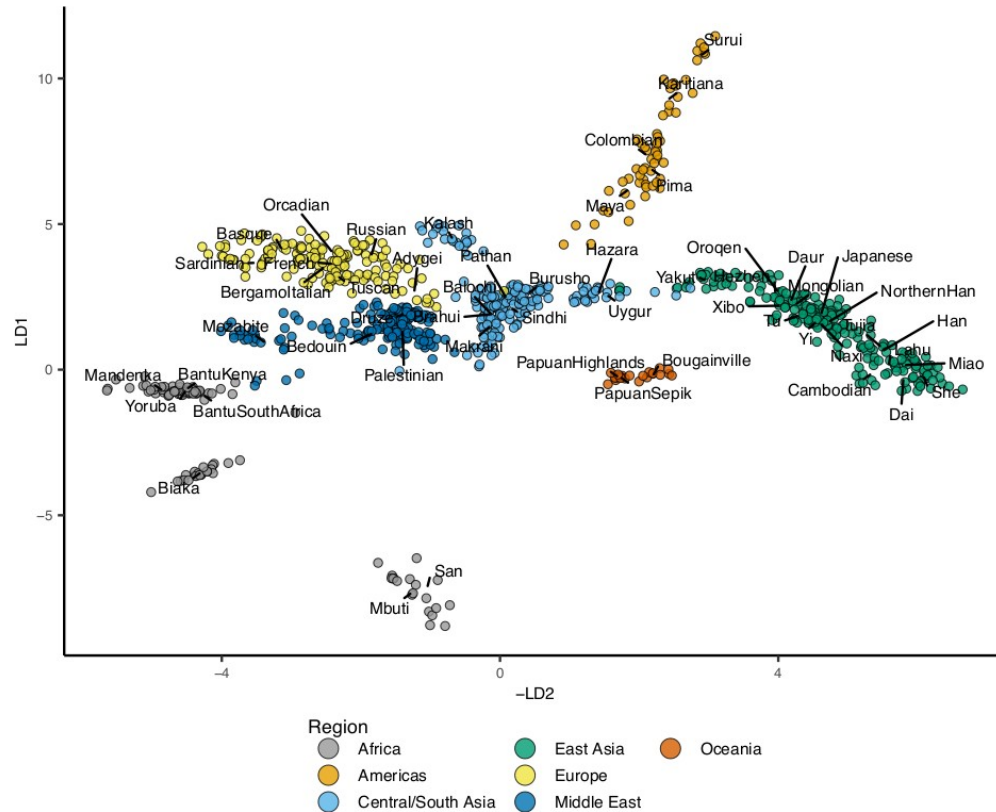
Battey, C. J., Gabrielle C. Coffing, and Andrew D. Kern. "Visualizing population structure with variational autoencoders." G3 11.1 (2021): 1-11.

ADMIXTURE

So, which one is better?

No figure legend on purpose :-)

ADMIXTURE

So, which one is better?
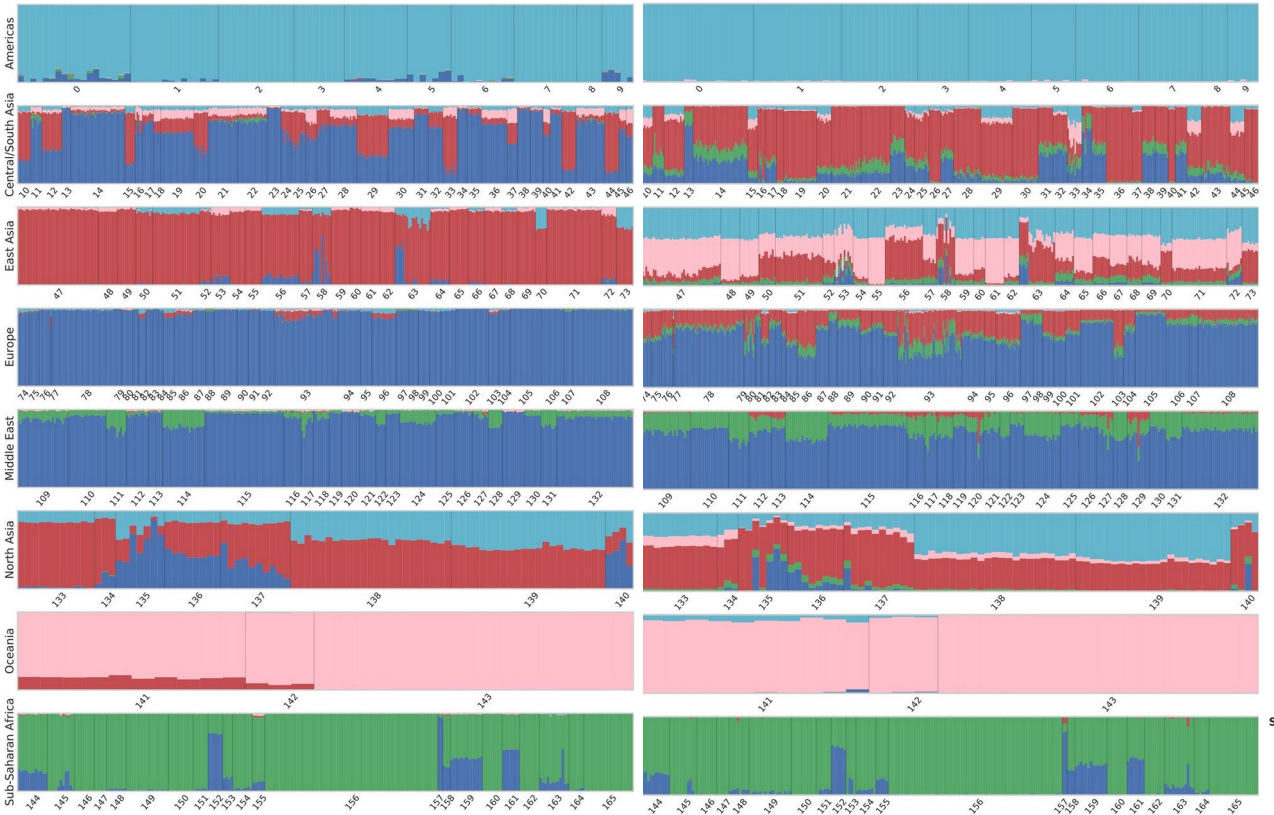
No figure legend on purpose :-)

Method

How is the ADMIXTURE-like plot generated?

Reproducible from text only?

No code

> [...] what you see is not
> ADMIXTURE w/PCA vs ADMIXTURE w/GenoCAE,
> but rather ADMIXTURE vs. k-pop softmaxed GenoCAE

F1 Scores
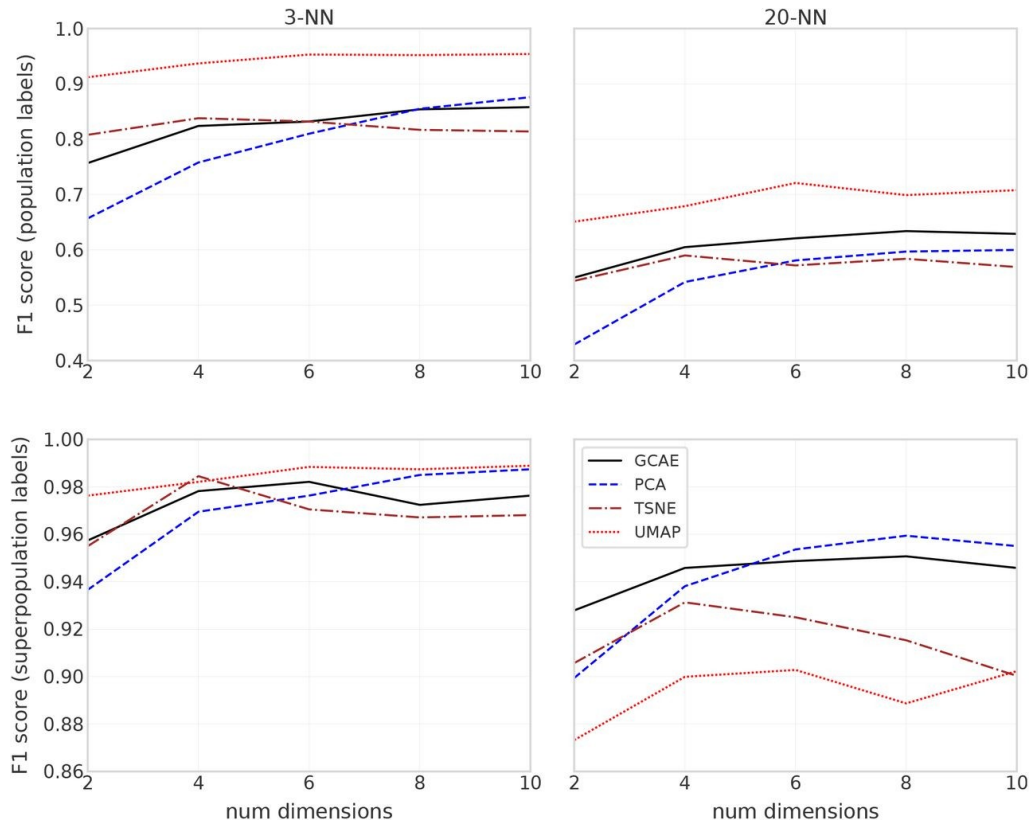
So, which one is better?

"The F1 score of a classification model based on the dimensionality reduction is not a simple metric for which the method with the highest score is the most correct"
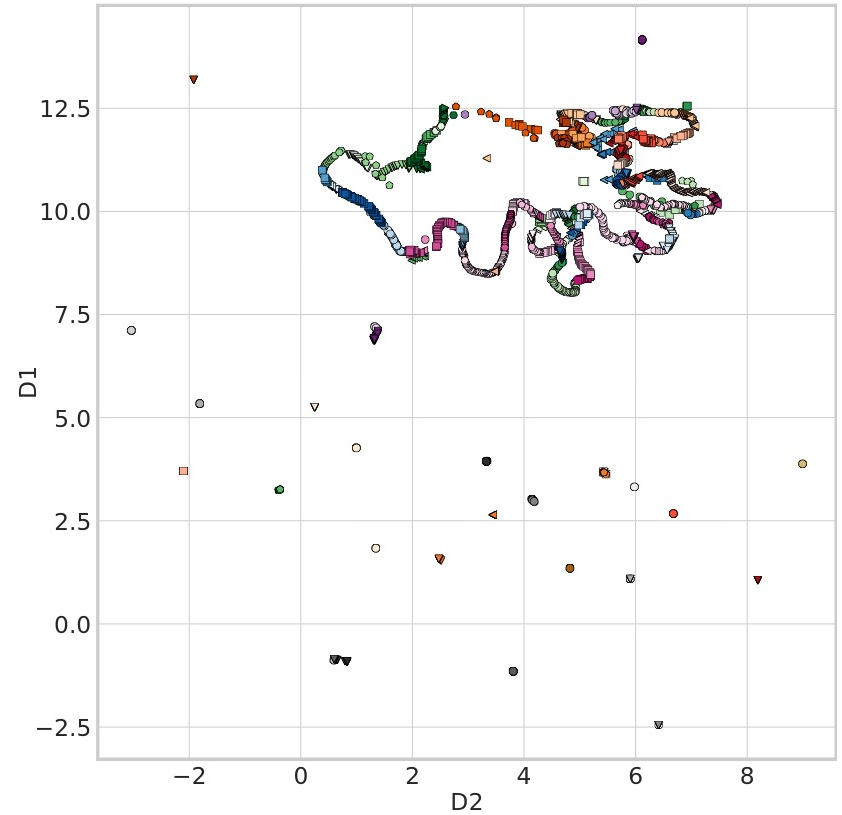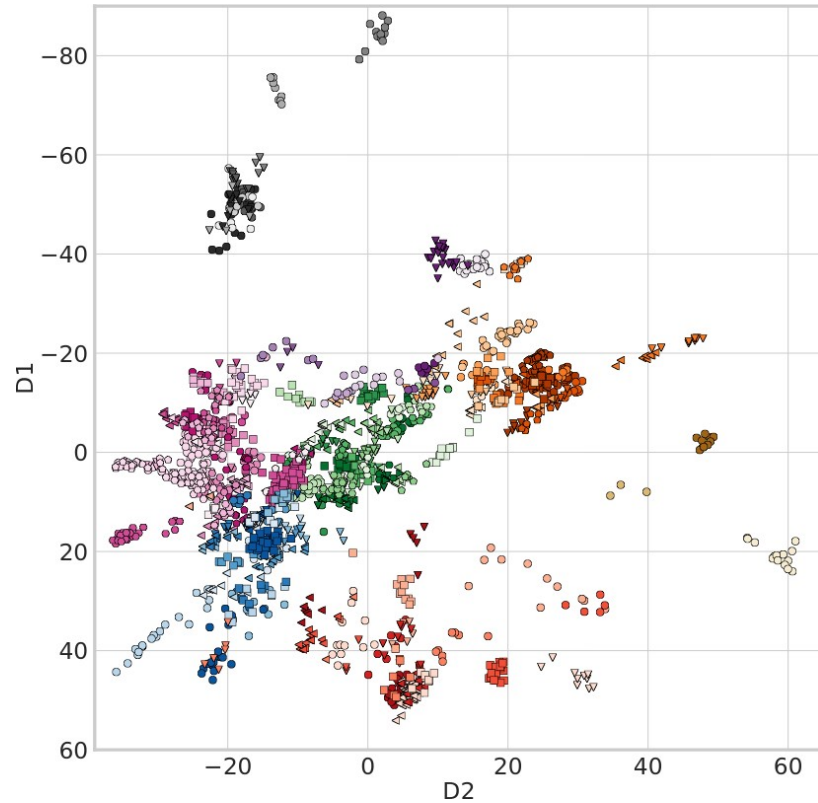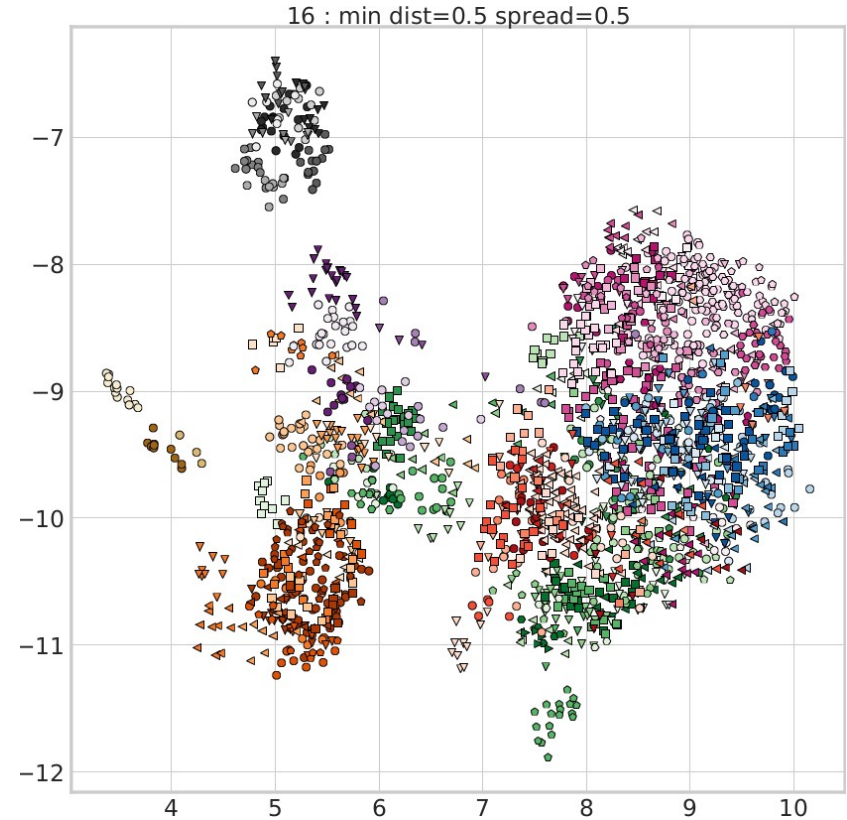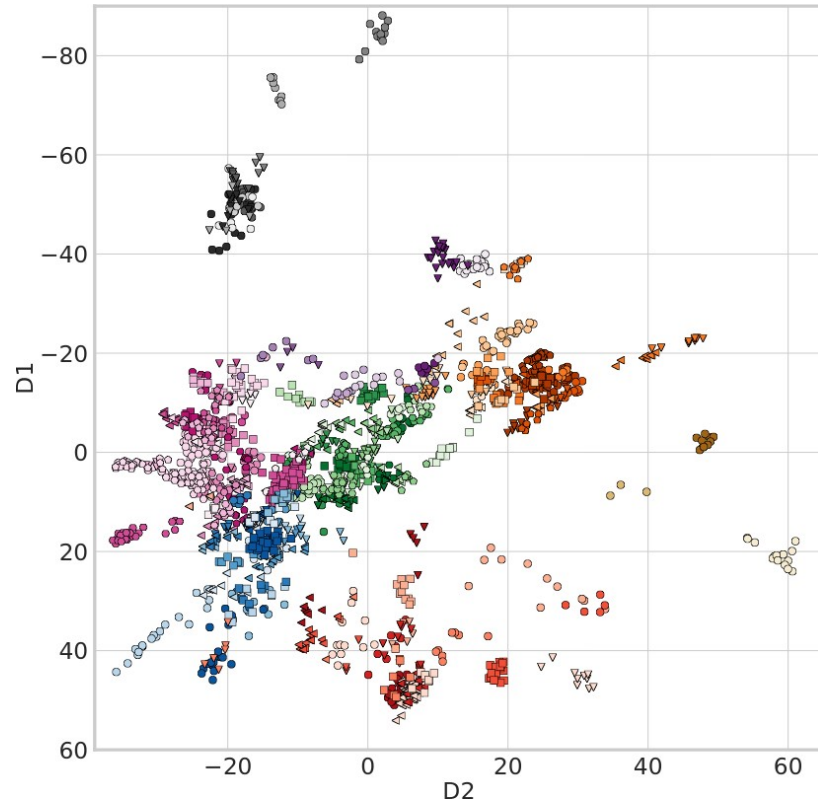
Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." bioRxiv (2020).

Discussion

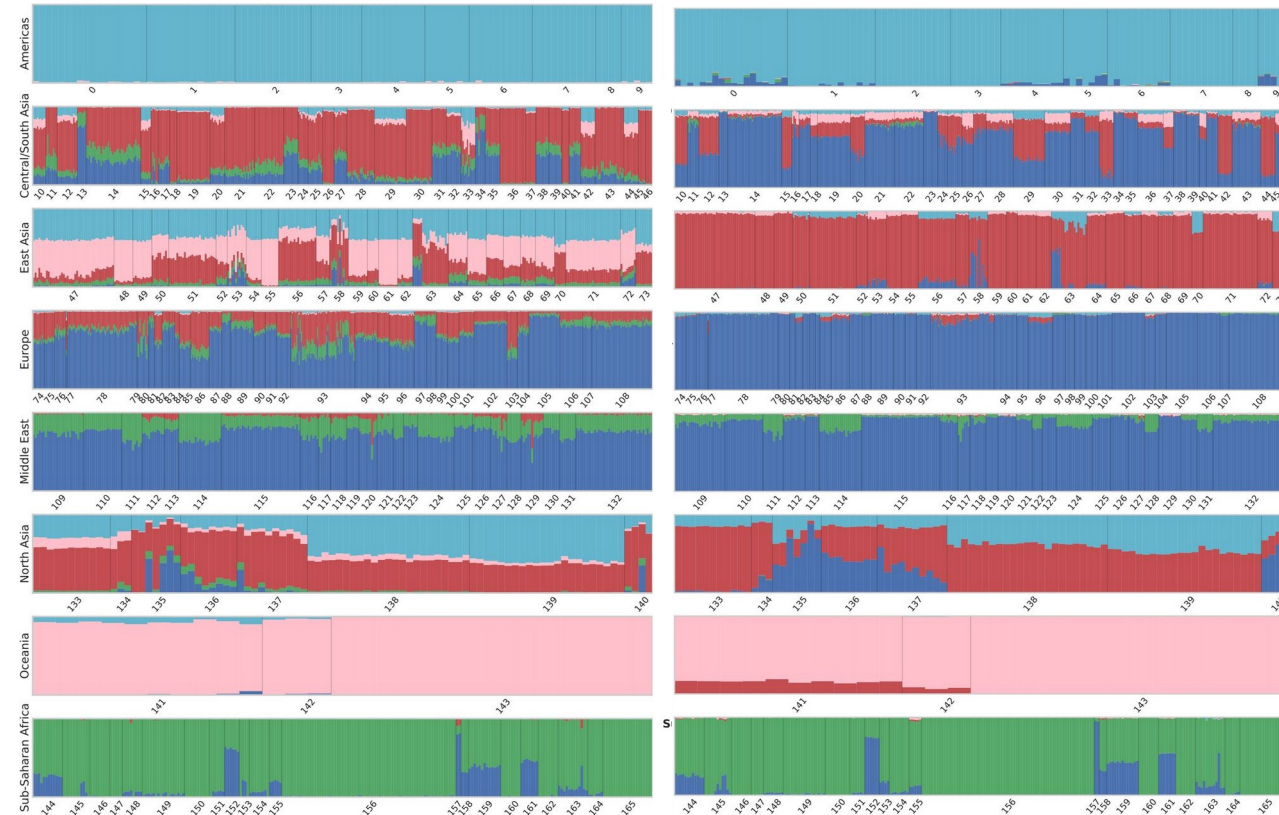**Novel approach**

… but there are others (non-convolutional) autoencoders

… best for non-linear effects on common SNPs

Dimensionality reduction looks impressive

… how well do other autoencoders do?

Performance is measured

… but how conclusively?

Performance is compared

… but how much care/tuning for these comparisons?

The end

https://github.com/richelbilderbeek/
journal_club_20220220

Data removal procedure

Human Origins data set

[...]

The data was filtered to exclude sex chromosomes and non-informative sites, and *one sample* (NA13619) was removed due to relation to another (HGDP01382).

Because of ADMIXTURE assumption

Equations

$$E(y, \hat{y}) = \boxed{\sum_{i}^{3} y_i log(\hat{y}_i)} + \alpha \sum_{j}^{d} e_j^2$$

$E(y, y^{hat})$: error

i: one of the three variants, i.e. AA, AC, CC

y_i: the actual value of a variant, i.e.

AA = 0.0, AC = 0.5, CC = 1.0

y^{hat}_i: the decoded value of a variant

```
"loss": {
  "module": "tf.keras.losses",
  "class": "CategoricalCrossentropy",
  "args": {
"from_logits": false}},
"regularizer": {
  "reg_factor": 1.0e-07,
  "module": "tf.keras.regularizers",
  "class": "l2"
},
```

Equations

$$E(y, \hat{y}) = \sum_{i}^{3} y_i log(\hat{y}_i) + \alpha \sum_{j}^{d} e_j^2$$

Lower = better

G	y_i	y^{hat}_i	E
AA	0	0.5	0
AC	1	0.1	-2.3
CC	0	0.3	0

One-hot encoding

Zero = best

Lower = better

Equations

$$E(y, \hat{y}) = \sum_{i}^{3} y_i \log(\hat{y}_i) + \alpha \sum_{j}^{d} e_j^2$$

i: one of the three variants, i.e. AA, AC, CC

y_i: the actual value of a variant, i.e.

AA = 0.0, AC = 0.5, CC = 1.0

y^{hat}_i: the decoded value of a variant

```
"loss": {
    "module": "tf.keras.losses",
    "class": "CategoricalCrossentropy",
    "args": {
"from_logits": false}},
"regularizer": {
    "reg_factor": 1.0e-07,
    "module": "tf.keras.regularizers",
    "class": "l2"
},
```