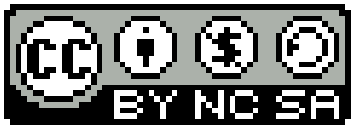
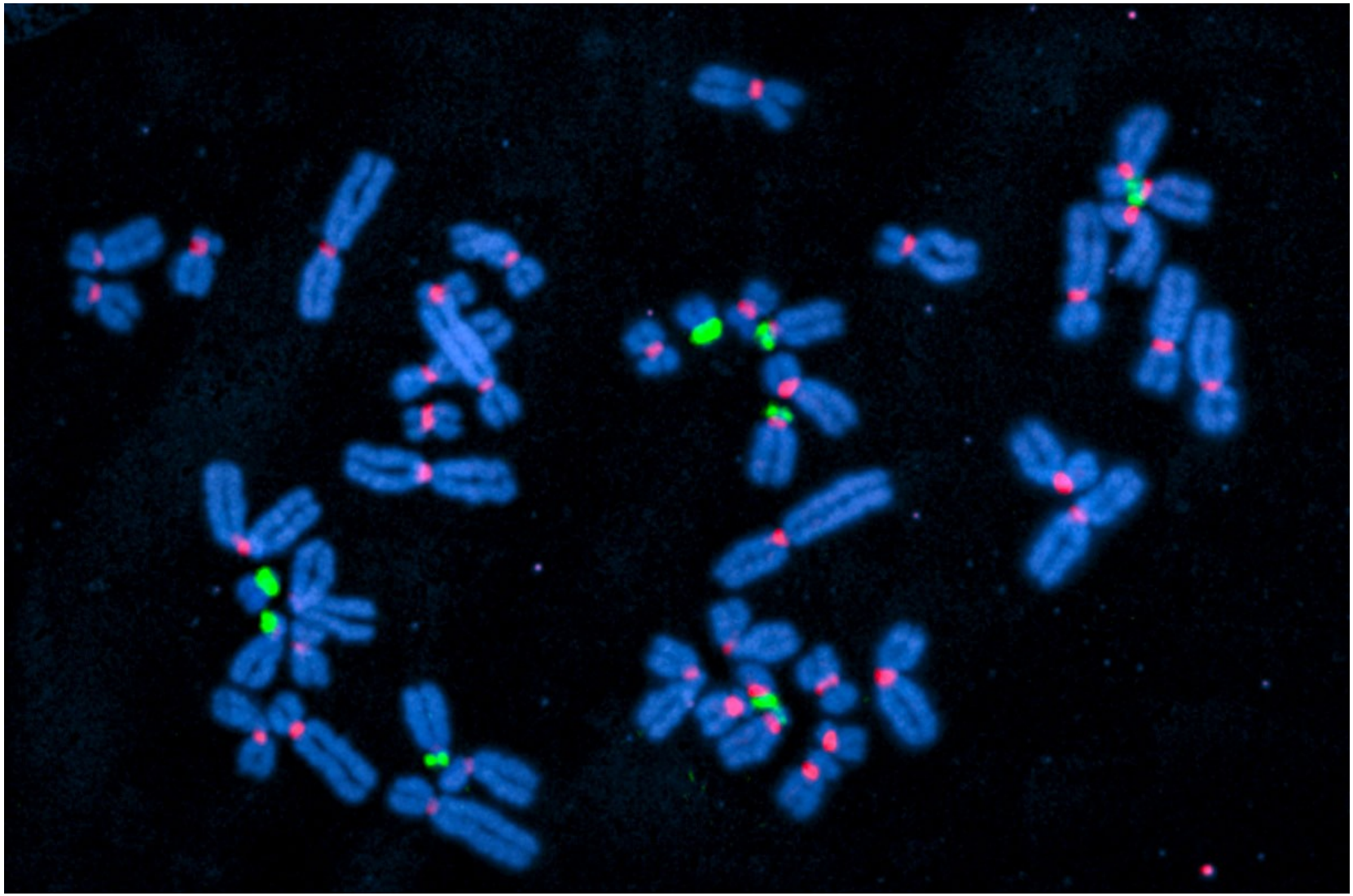
Quantitative trait prediction using GCAE

Richel Bilderbeek
2022-03-05



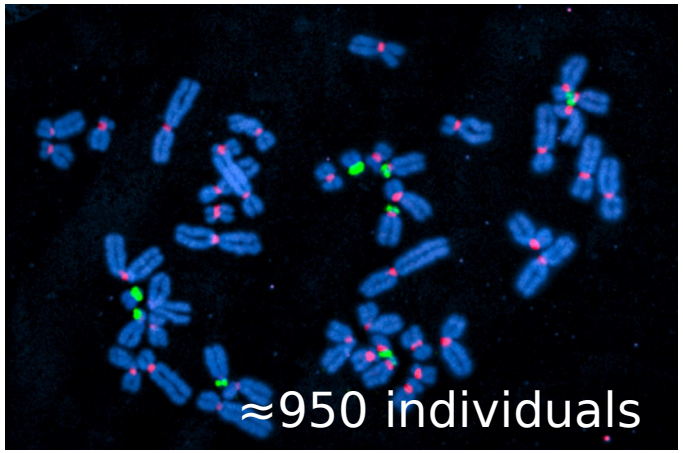
https://github.com/richelbilderbeek/science_presentation_20220305

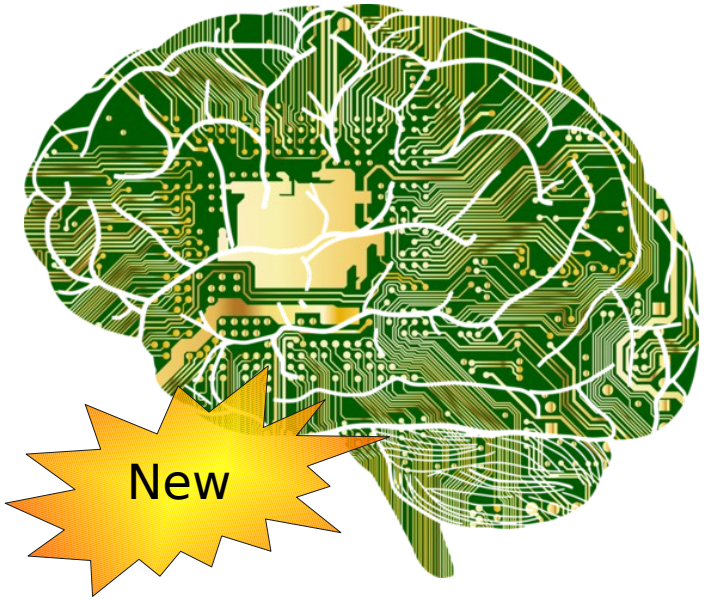
Introduction



Picture from Elizabeth Pennisi. "Most complete human genome yet reveals previously indecipherable DNA". Science (2022)
Deanna M. Church. "A next-generation human genome sequence." Science 376: (2022): 34-35. DOI: 10.1126/science.abo5367

Goal





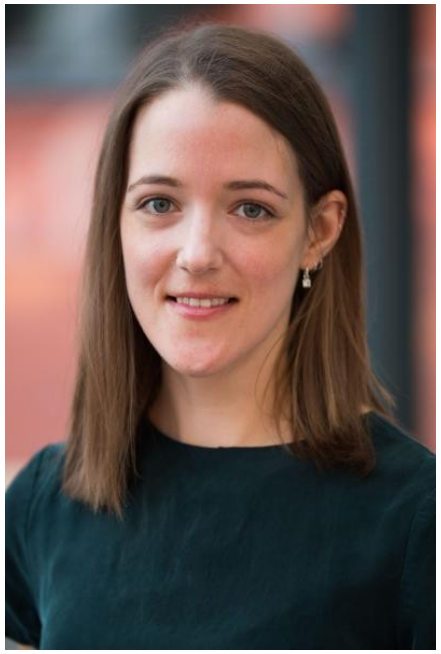
Predict

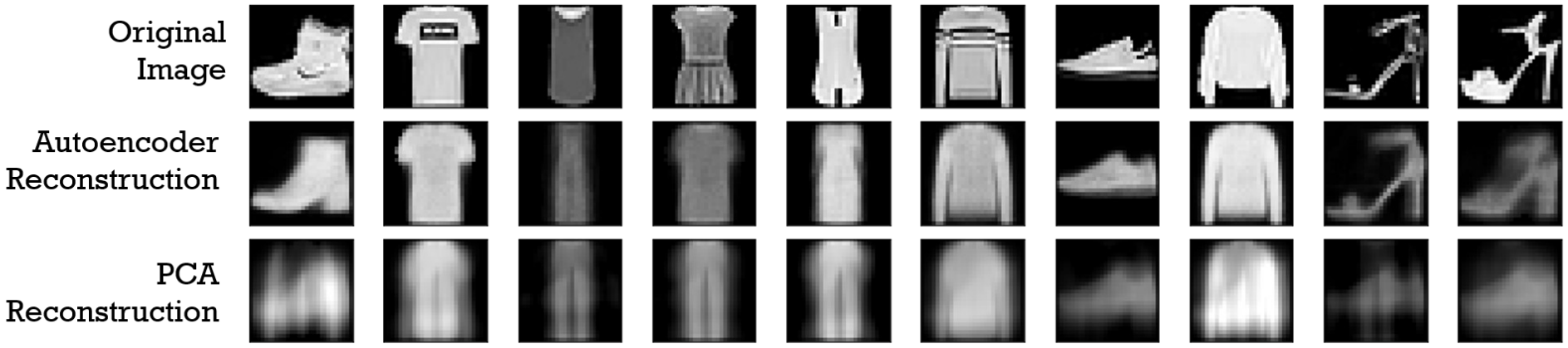
Elizabeth Pennisi. "Most complete human genome yet reveals previously indecipherable DNA". Science (2022)
https://en.wikipedia.org/wiki/File:Anatomy-1751201_1280.png
https://commons.wikimedia.org/wiki/Category:Phenotypes#/media/File:Coquina_variation3.jpg
https://commons.wikimedia.org/wiki/File:Blood-tube-fig-2.png

Method: GCAE

Genomic Convolutional Auto Encoder

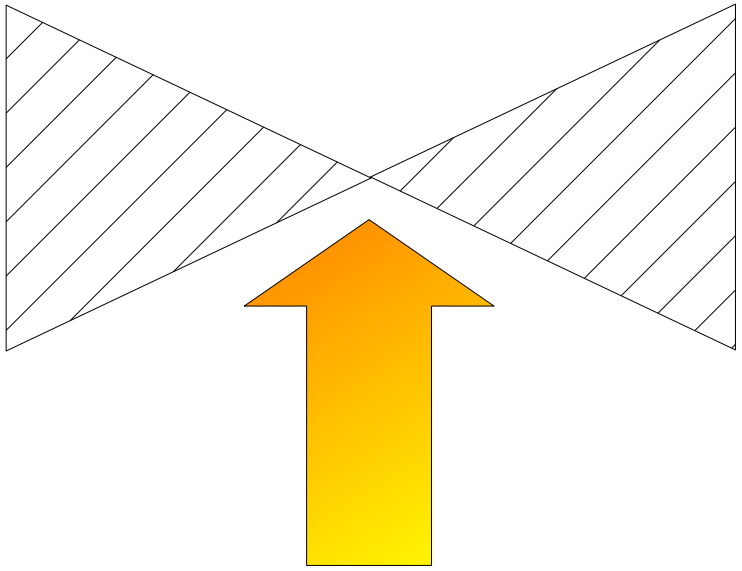
Dimensionality reduction

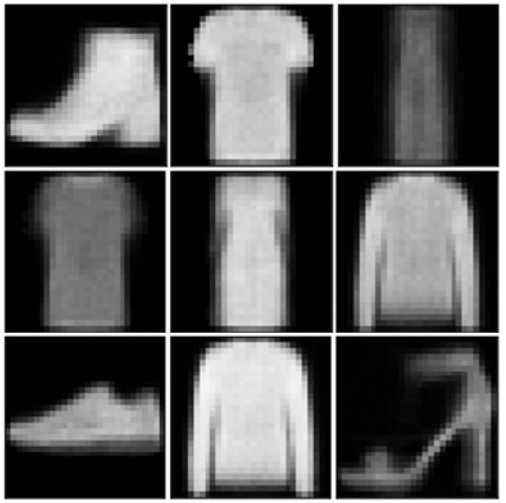




Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." G3 12.3 (2022): jkac020. https://doi.org/10.1093/g3journal/jkac020
https://en.wikipedia.org/wiki/File:Reconstruction_autoencoders_vs_PCA.png

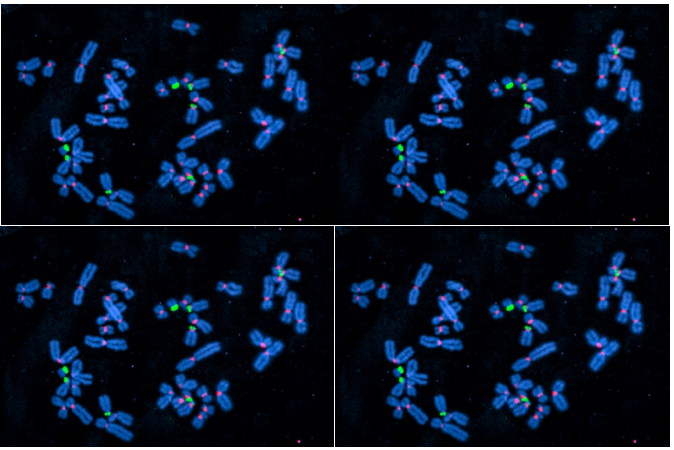
What an auto-encoder does

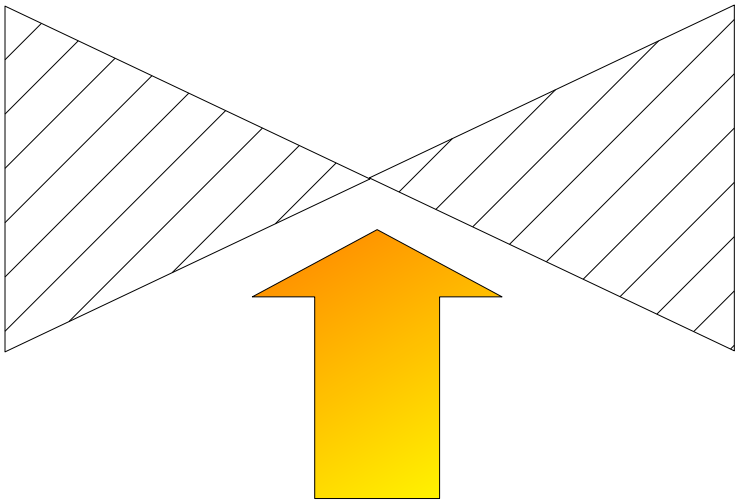


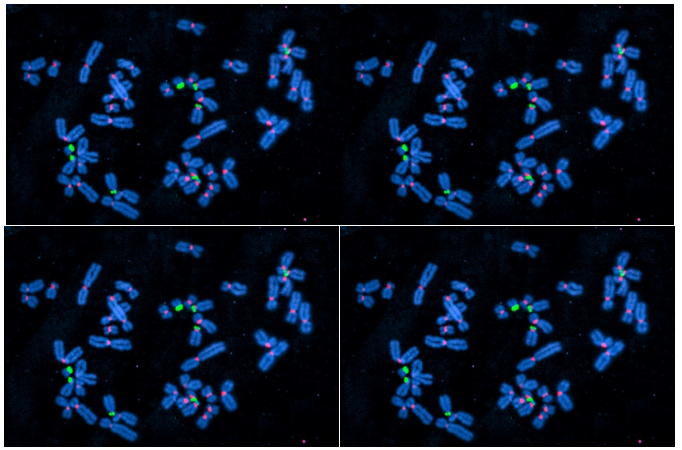


Latent layer with dimension reduced data

GCAE is an auto-encoder

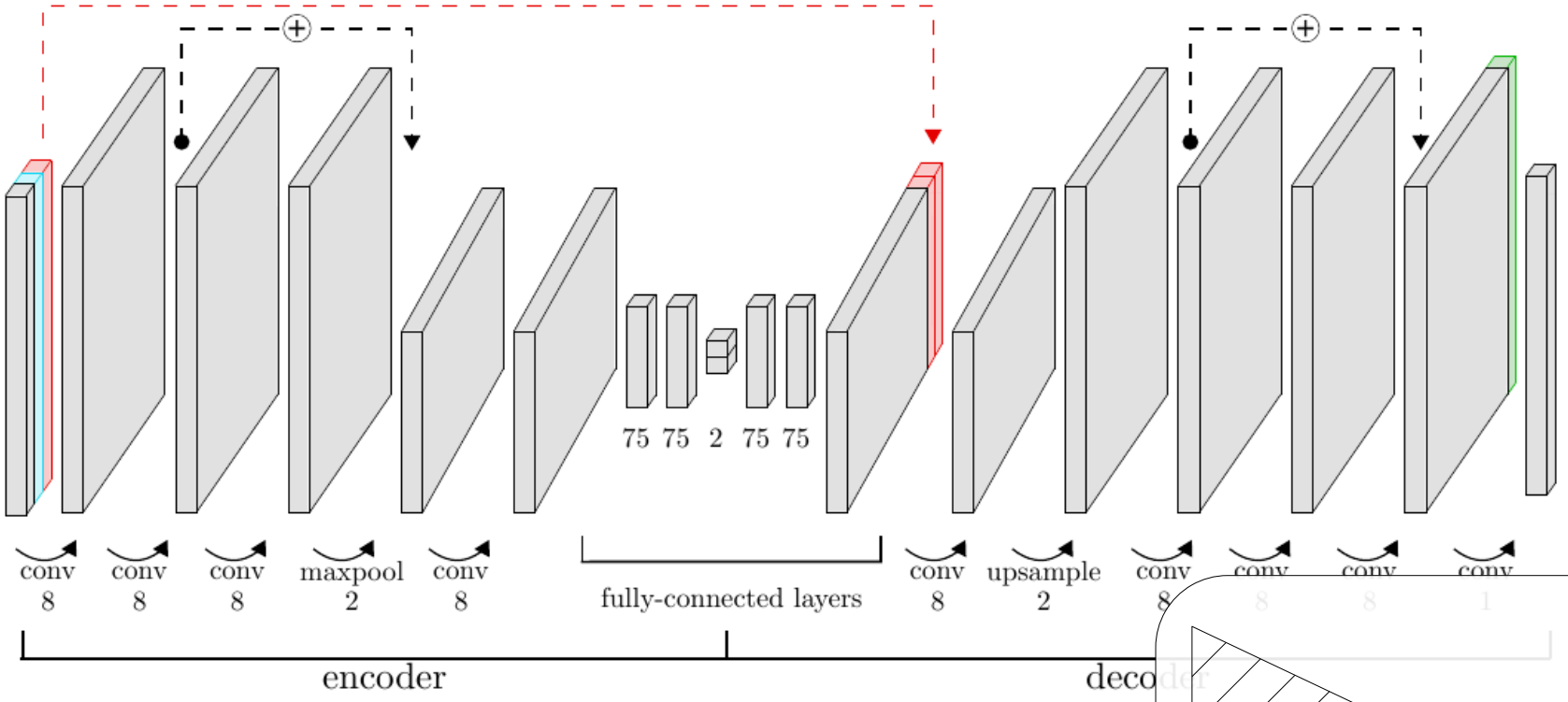


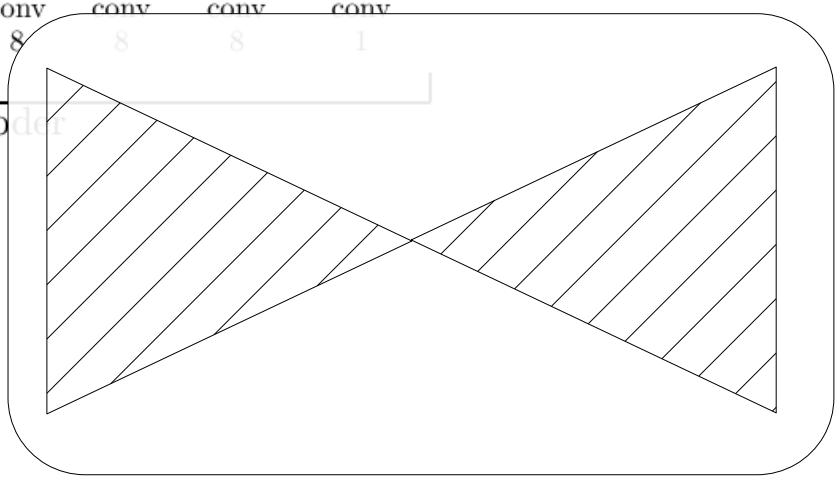




https://commons.wikimedia.org/wiki/File:Disorganized_books_piling_one_another.JPG2
https://commons.wikimedia.org/wiki/File:Stadsbiblioteket_i_Uppsala_interior_04.jpg

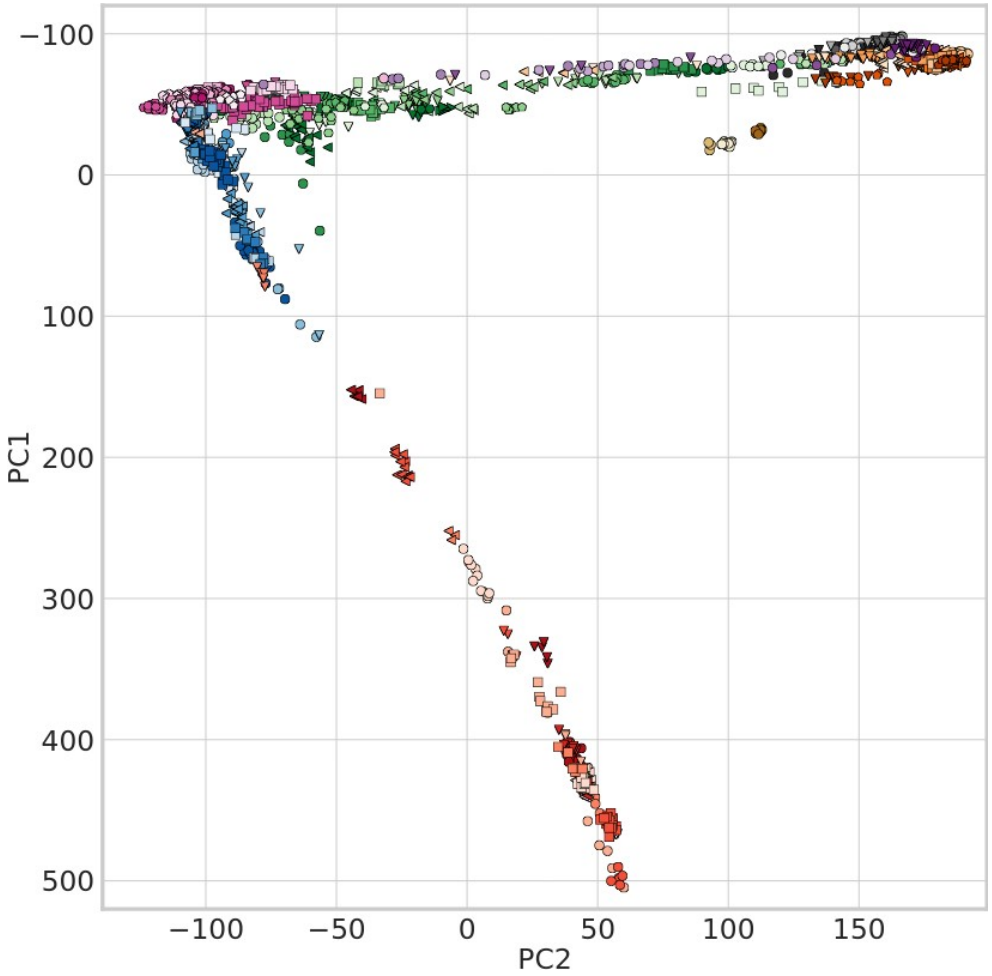
GCAE is highly tunable

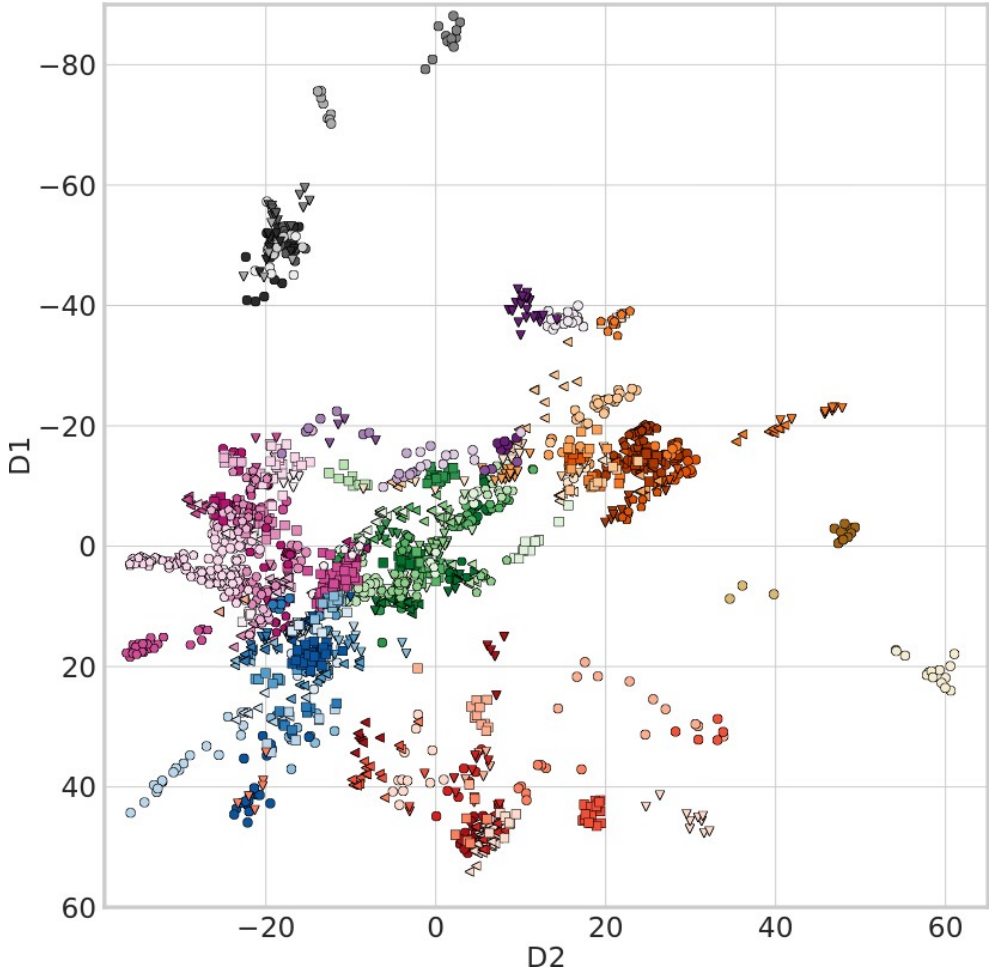




https://github.com/kausmees/GenoCAE/blob/master/images/example_architecture.png

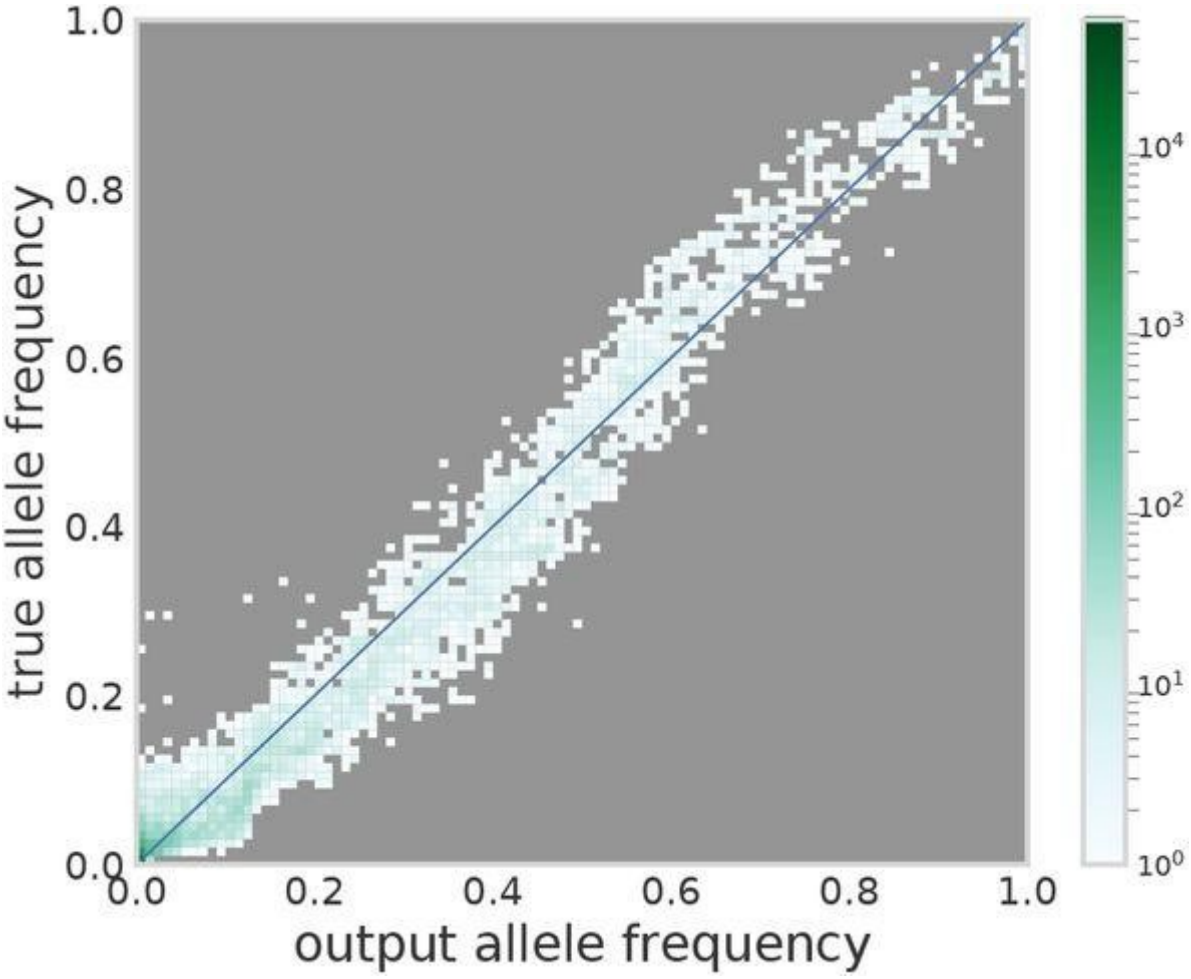
GCAE can do non-linear dimensionality reduction

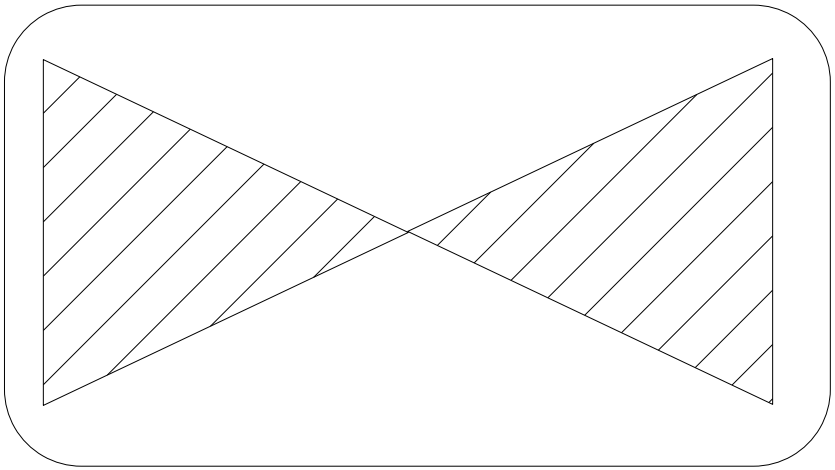




Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." G3 12.3 (2022): jkac020. https://doi.org/10.1093/g3journal/jkac020

GCAE cannot do rare alleles

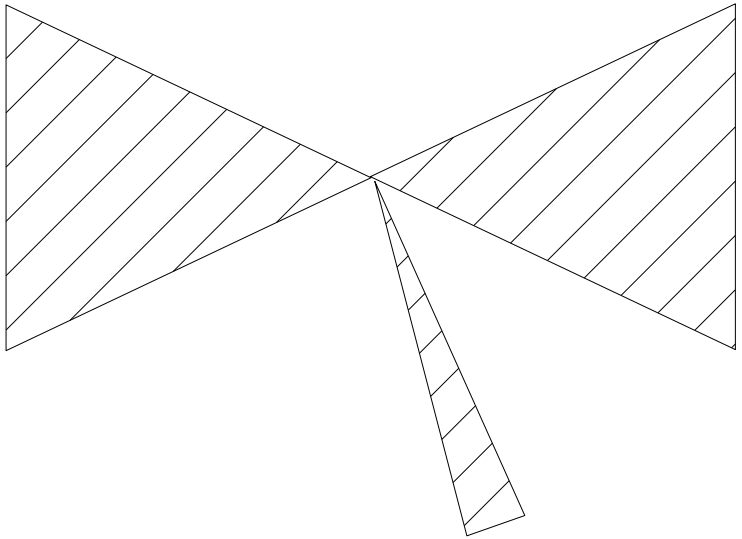




Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." G3 12.3 (2022): jkac020. https://doi.org/10.1093/g3journal/jkac020

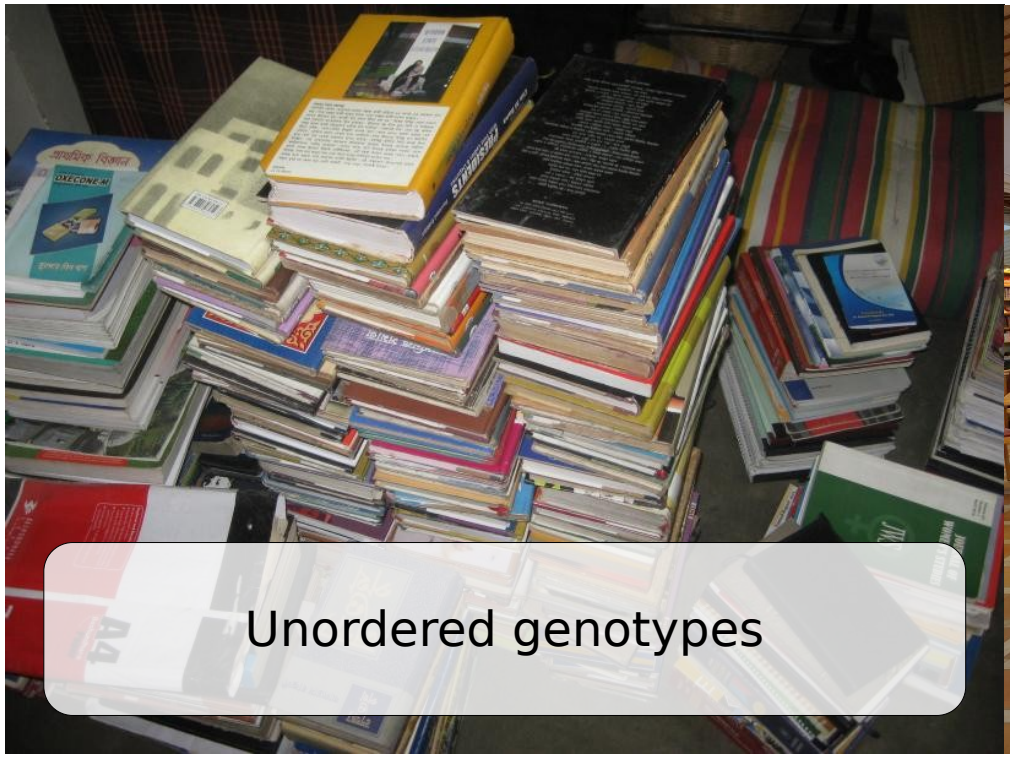
GCAE may do trait prediction

Untested quantitative trait prediction extension

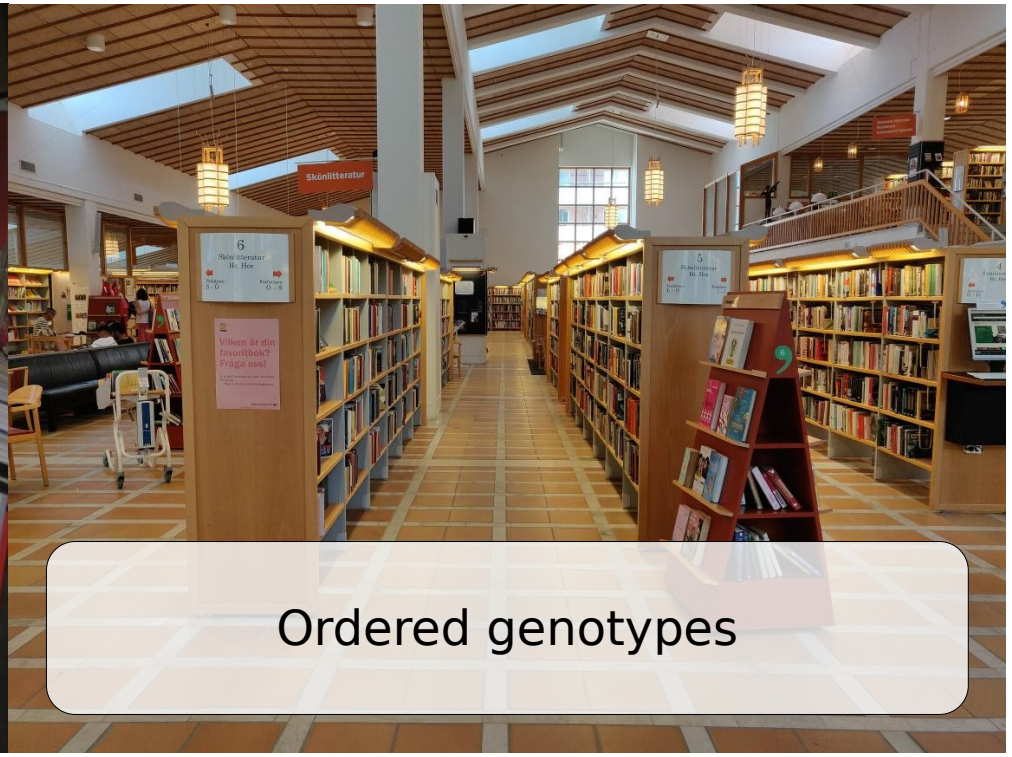


New

Analogy



Unordered genotypes



Ordered genotypes

Predict the number of horses, spaceships, etc.

Protein concentrations

https://commons.wikimedia.org/wiki/File:Disorganized_books_piling_one_another.JPG2

https://commons.wikimedia.org/wiki/File:Stadsbiblioteket_i_Uppsala_interior_04.jpg

What has been done

GCAE

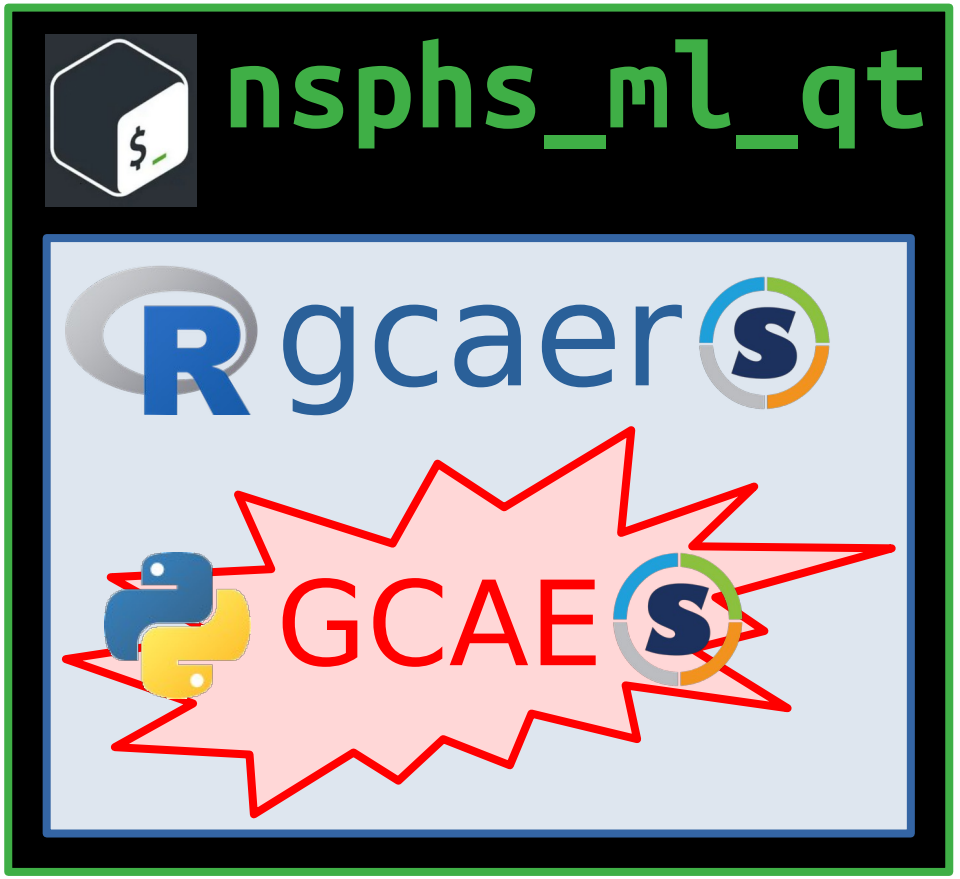
- Can actually run with trait prediction
- Can run on Bianca

gcaer works with GCAE

- Tested to be correct
- Proper error messages
- Extend functionality

nsphs_ml_qt does experiments

- Scripts re-used in many contexts
- Works locally, on Rackham, on Bianca
- You can run these today



https://github.com/richelbilderbeek/GenoCAE/tree/Pheno
https://github.com/richelbilderbeek/gcaer
https://github.com/richelbilderbeek/nsphs_ml_qt

First experiment

Simulate simplest dataset

Run GCAE as-is

Evaluate performance

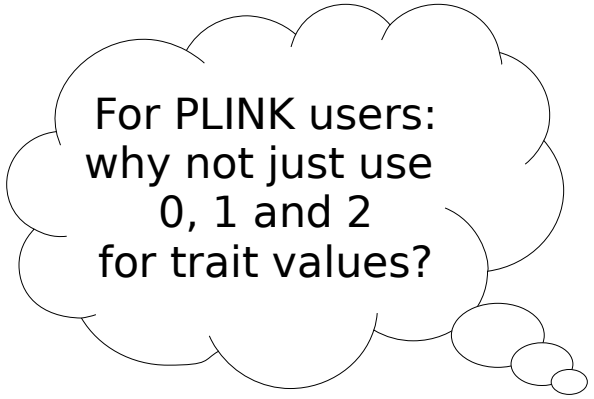
https://github.com/richelbilderbeek/nsphs_ml_qt/scripts_rackham/02_start_2_richel_issue_126.sh

First experiment: simulated dataset

1 monogenic trait with $H^2 = h^2 = 1$

1k individuals

MAF = 0.499



chr	id	posg	pos	ref	alt
1	snp_1	0	1	A	C

FID	IID	additive
A	1	9.424778
A	2	6.283185
A	3	3.141593
A	4	9.424778
A	5	9.424778

	snp_1
1	0
2	1
3	2
4	0
5	0

fam	id	pat	mat	sex	pheno
A	1	0	0	1	-9
A	2	0	0	1	-9
A	3	0	0	1	-9
A	4	0	0	1	-9
A	5	0	0	1	-9

https://github.com/richelbilderbeek/science_presentation_20220305/blob/master/analysis_richel_126.R

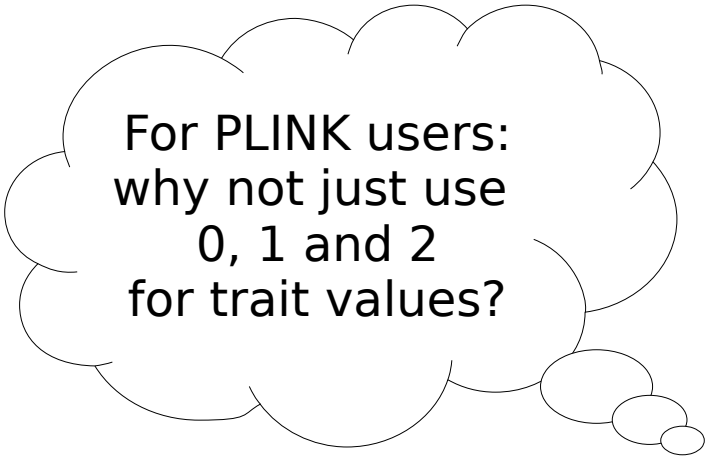
First experiment: PLINK results

PLINK measures association perfectly

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
1	snp_1	1	1000	-3.142	0	1	-116700000	0

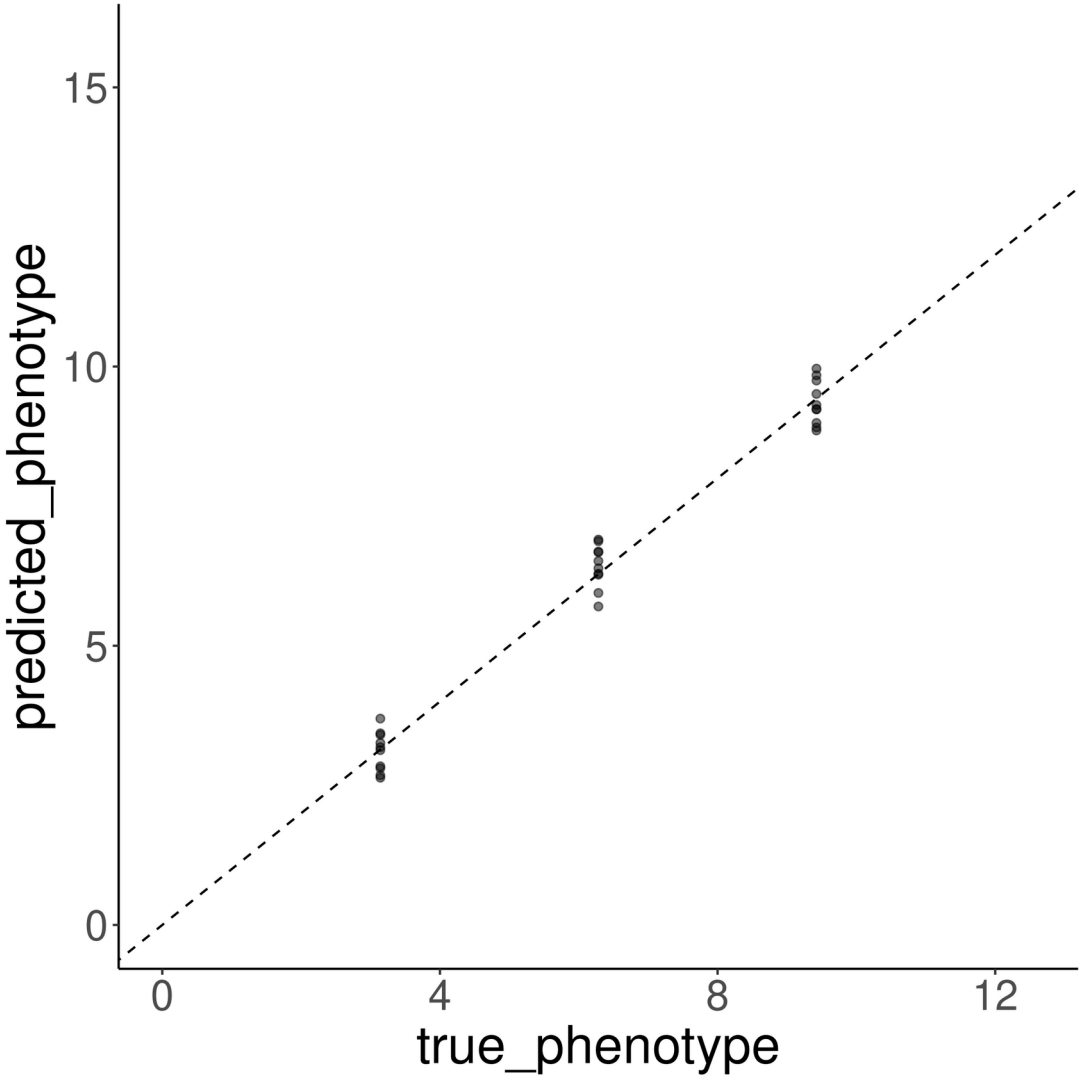
FID	IID	additive
A	1	9.424778
A	2	6.283185
A	3	3.141593
A	4	9.424778
A	5	9.424778

	snp_1
1	0
2	1
3	2
4	0
5	0



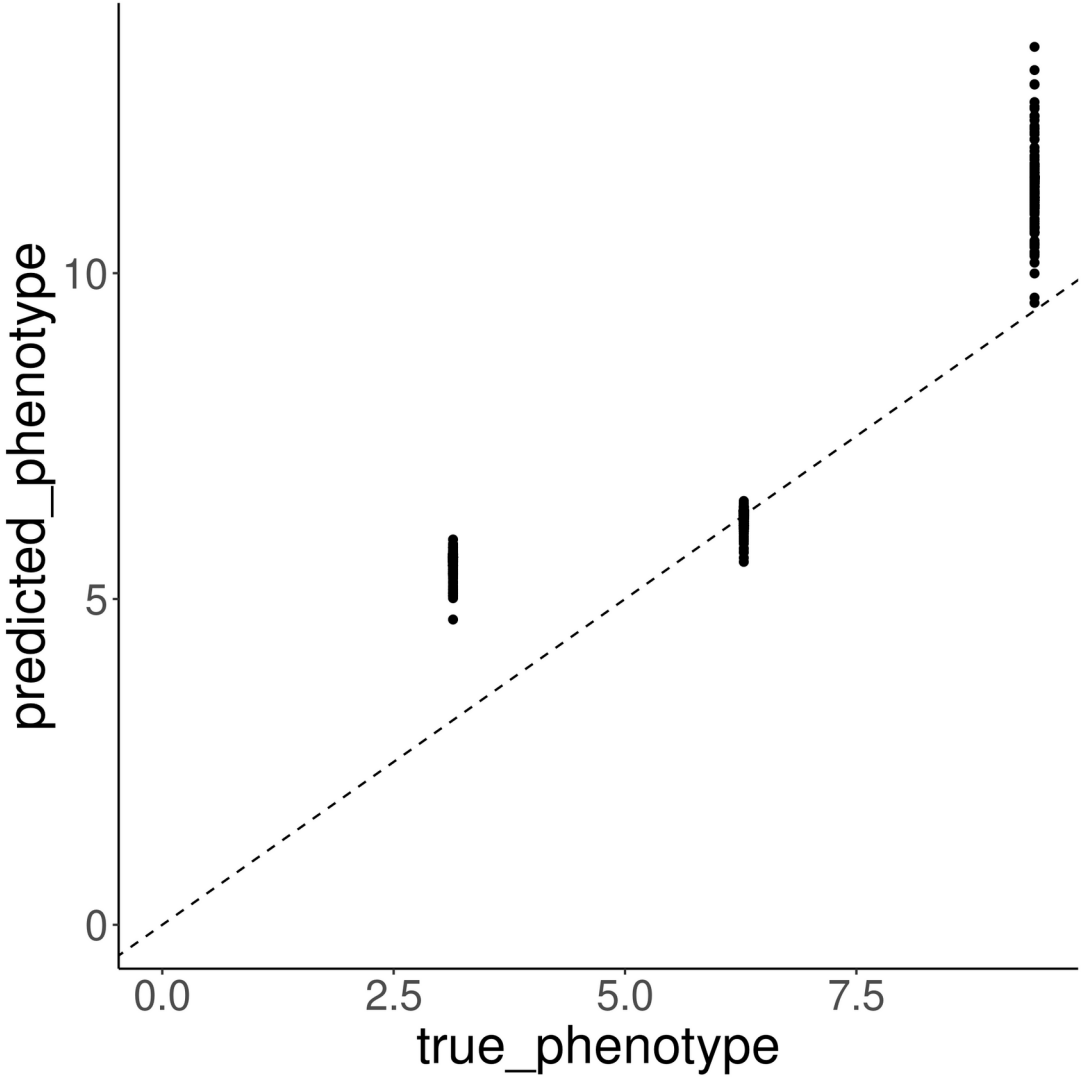
https://github.com/richelbilderbeek/science_presentation_20220305/blob/master/analysis_richel_126.R

First experiment: expected GCAE predictions



https://github.com/richelbilderbeek/science_presentation_20220305/blob/master/analysis_richel_126.R

First experiment: GCAE predictions

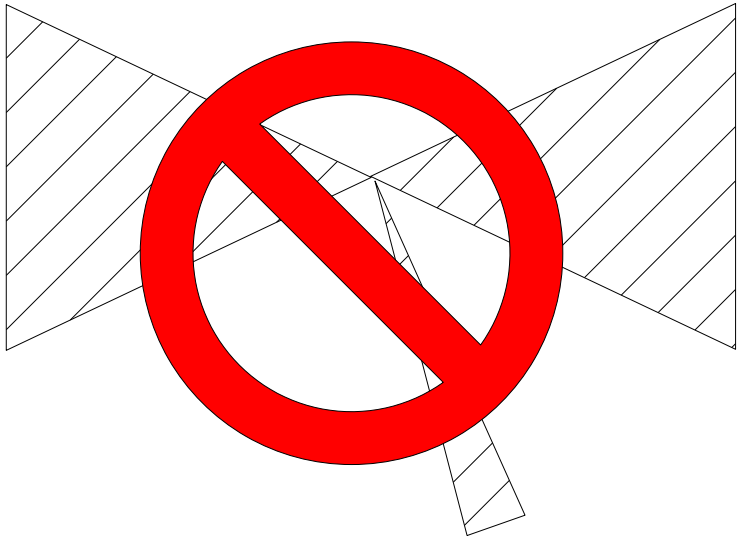


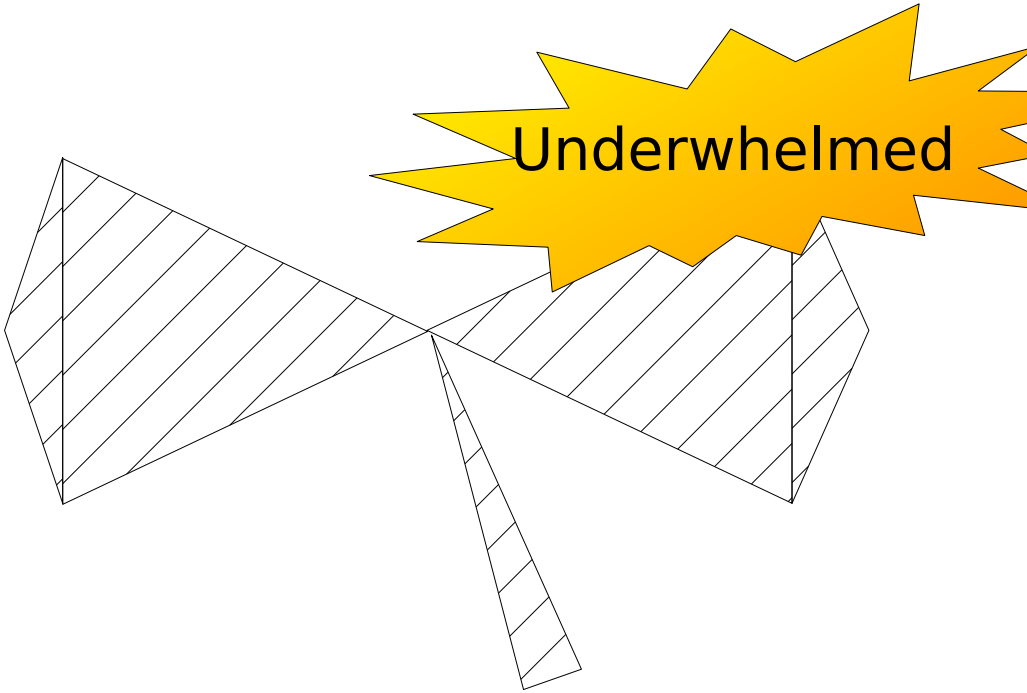
https://github.com/richelbilderbeek/science_presentation_20220305/blob/master/analysis_richel_126.R

First experiment: conclusions

Proof of concept works, prediction underperforms

Hypothesis:





Decision

Tune the network to the problem

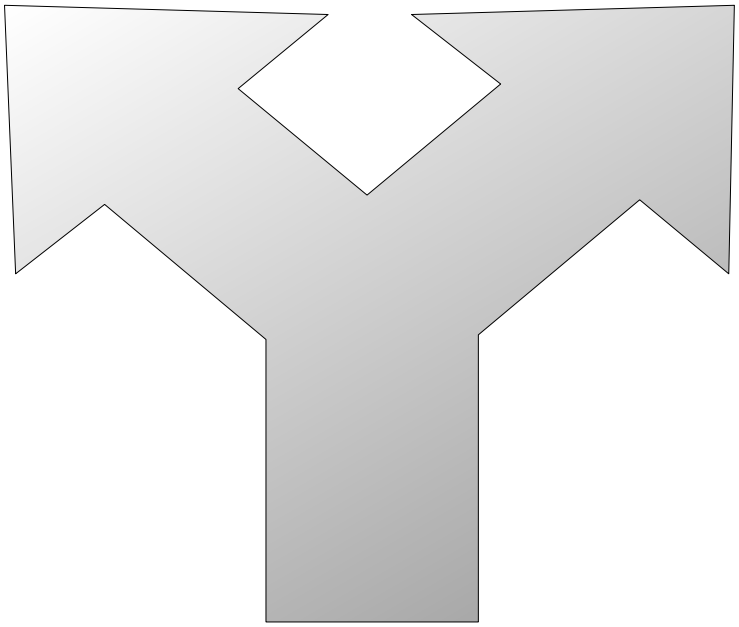
try a simpler neural network

GCAE focused

Tune the problem to the network

try harder problems

application focused



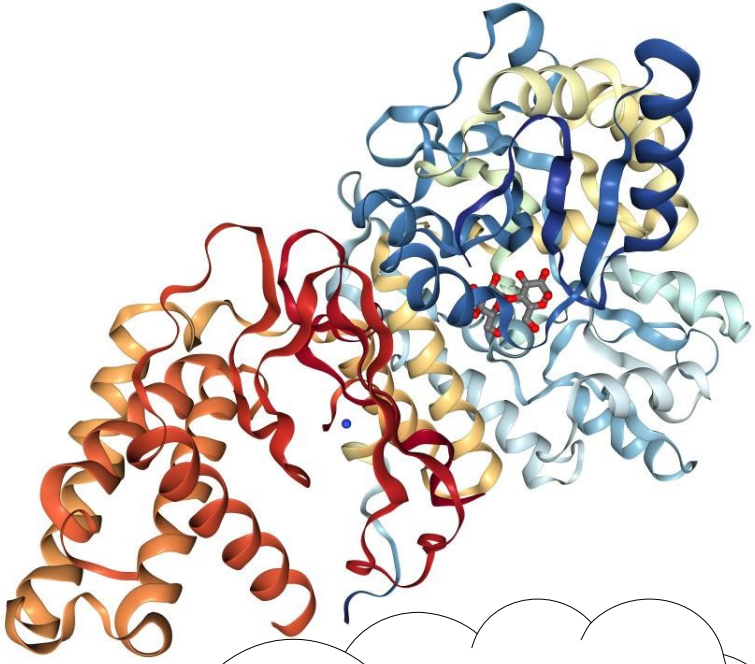
First NSPHS experiment

Use NSPHS and first protein concentration

- Pick 100k random variants
- Adrenomedullin

Run GCAE as-is

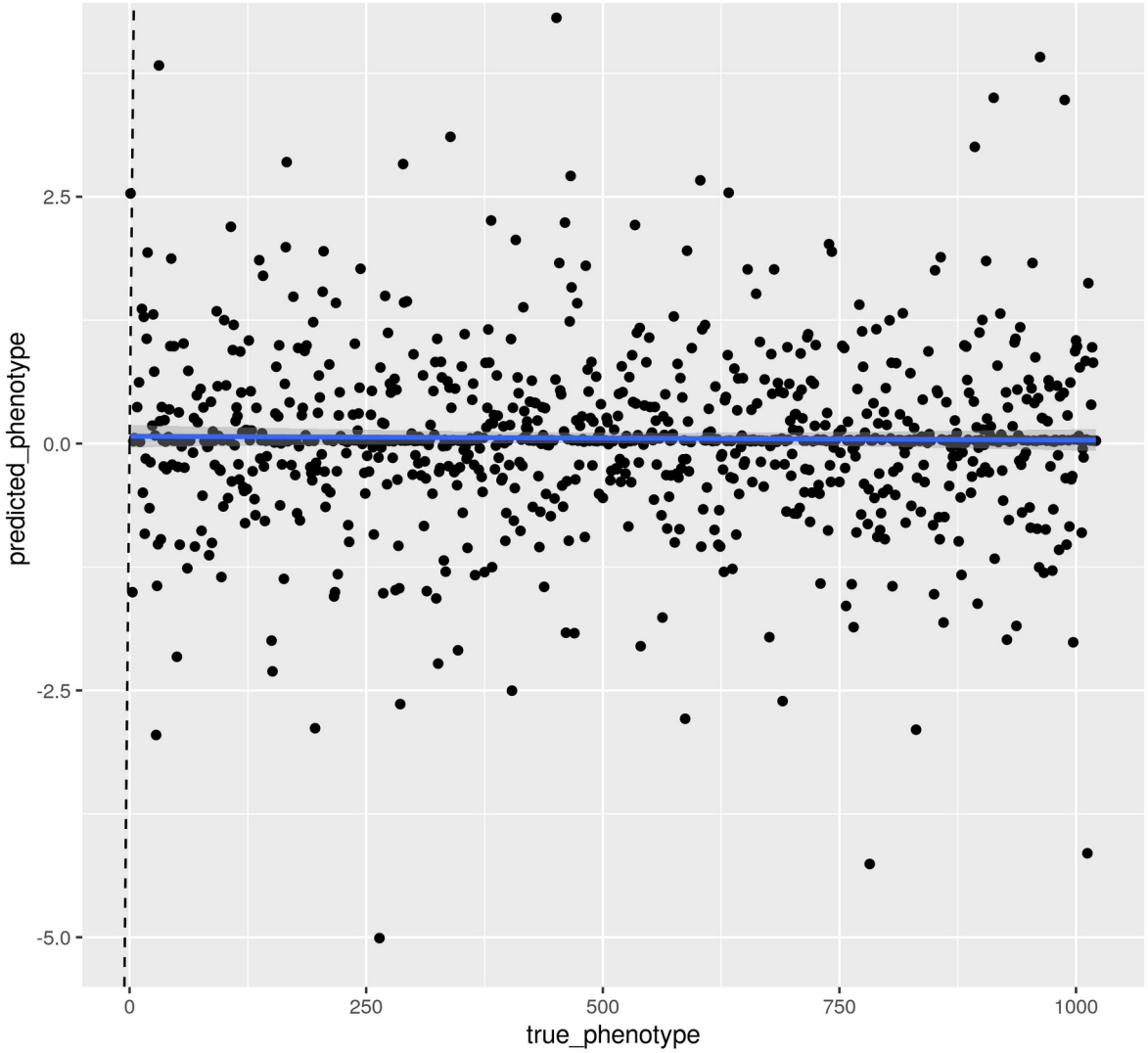
Evaluate performance



For NSPHS users: Can you predict the performance already?

https://cn.sinobiological.com/resource/adrenomedullin/proteins

First NSPHS experiment: results



Runs!

60 hours

1k epochs

Unsure about learning trajectory

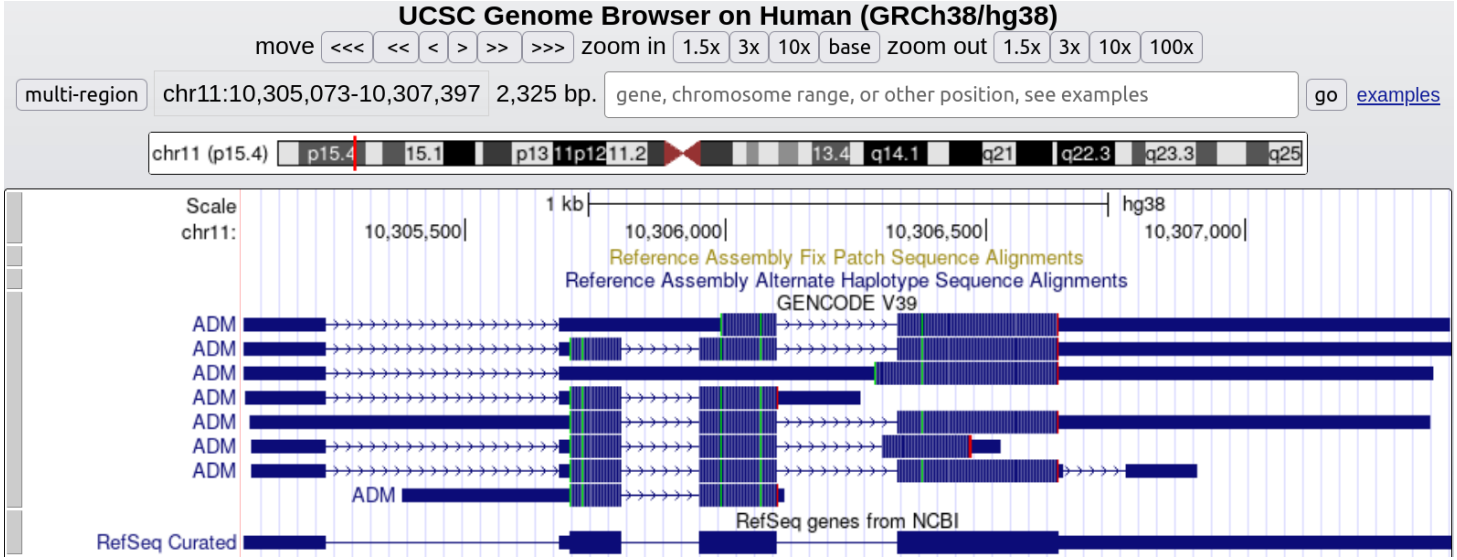
https://github.com/richelbilderbeek/nsphs_ml_qt/issues/14
https://github.com/richelbilderbeek/nsphs_ml_qt/blob/master/scripts_bianca/02_start_richel_issue_130.sh

First NSPHS: conclusion

The result is just noise

The gene for adrenomedullin is monogenic, 52 amino acids

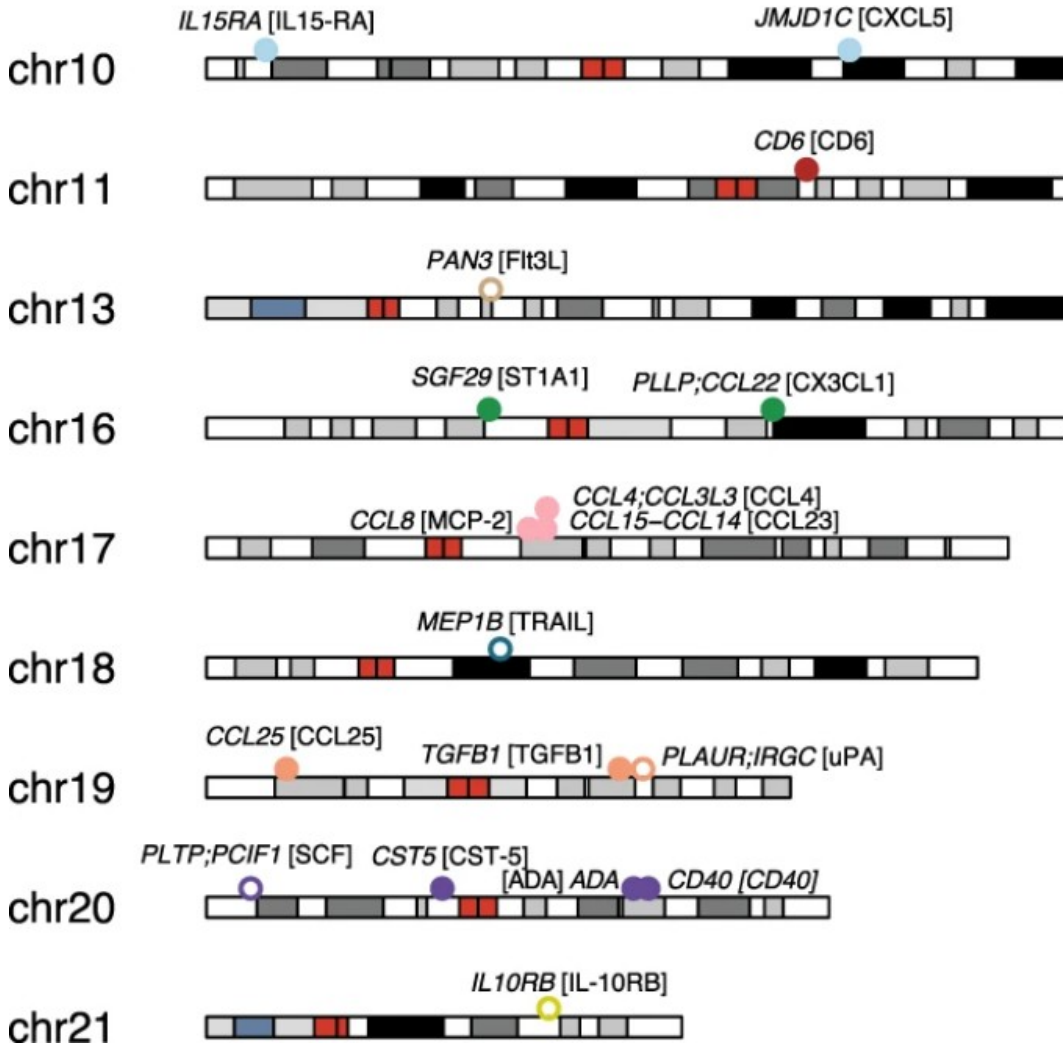
But hey, it works!



https://genome.ucsc.edu/cgi-bin/hgTracks?org=Human&db=hg38&position=chr11:10305073-10307397

First NSPHS: conclusion

* [ ] Use the SNPs known to have an association, e.g. 0.5 Mb downstream and upstream around known cis-regulatory elements, #5



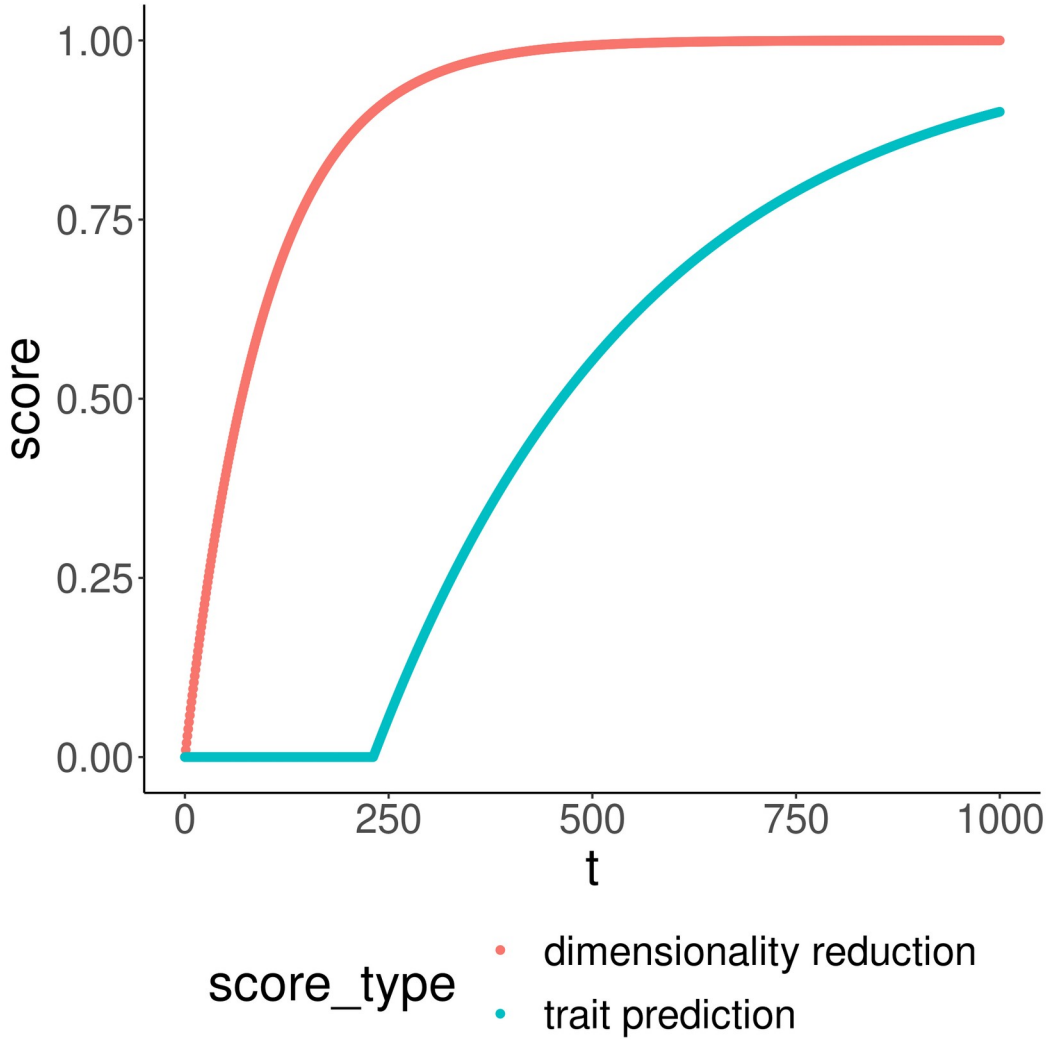
Höglund, Julia, et al. "Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers." Scientific reports 9.1 (2019): 1-14.
https://github.com/richelbilderbeek/nsphs_ml_qt/issues/5

Learning

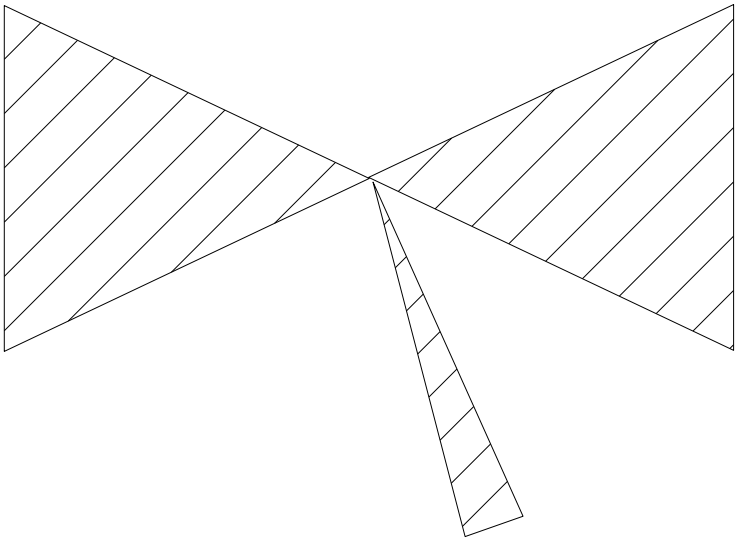
Learning proceeds
until death
and only then
does it stop

https://en.wikipedia.org/wiki/File:Portrait_of_Xun_Zi.jpg

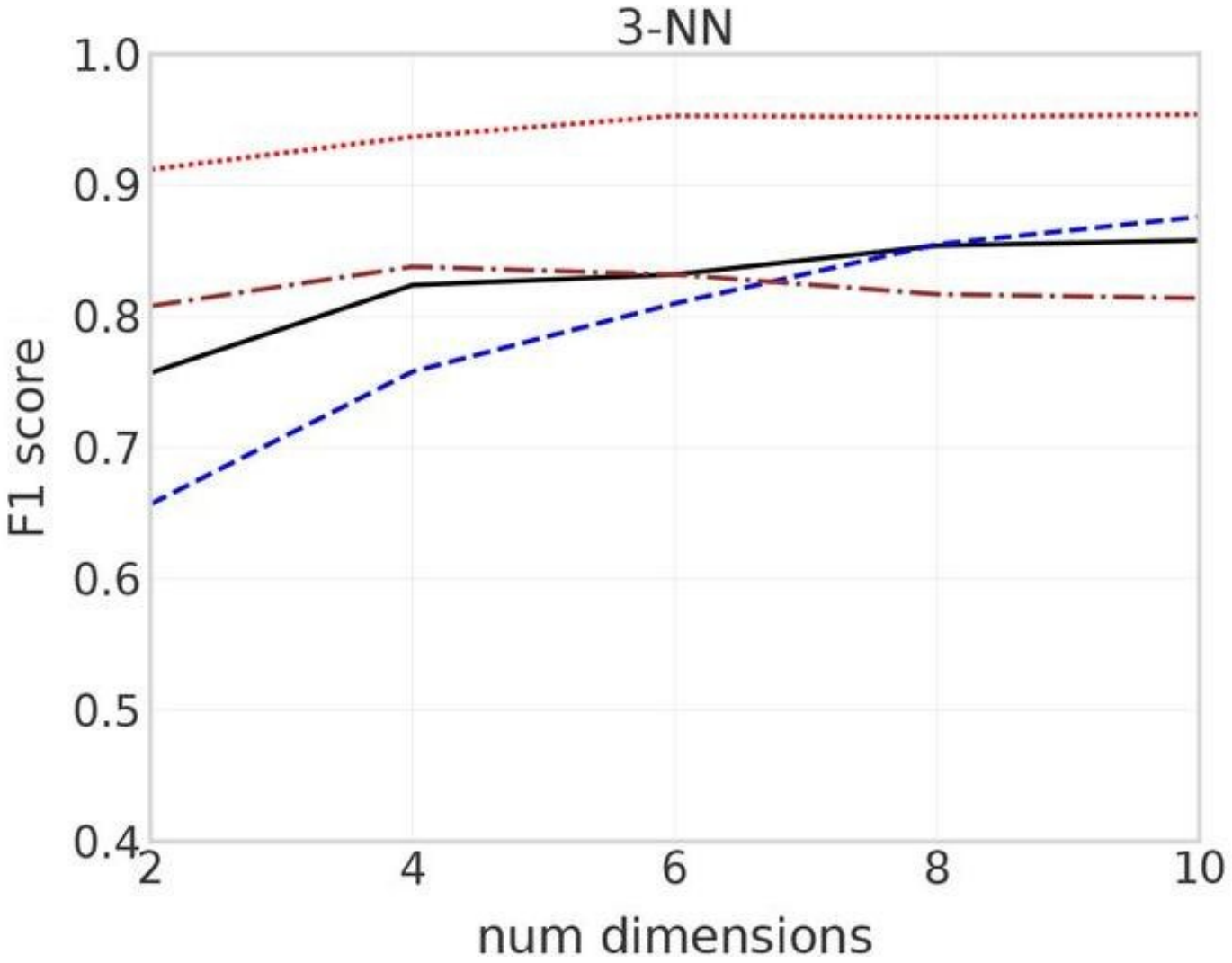
Learning trajectory

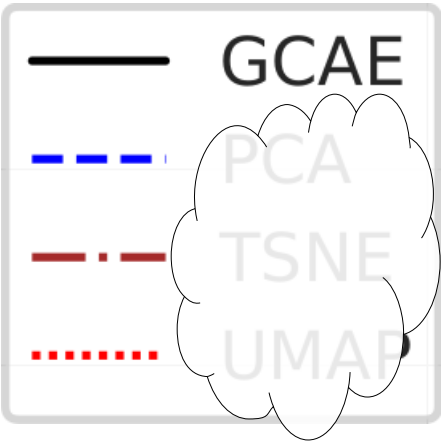


Trait prediction can only be trained well after dimensionality reduction is good



Dimensionality reduction score





Ausmees, Kristiina, and Carl Nettelblad. "A deep learning framework for characterization of genotype data." G3 12.3 (2022): jkac020. https://doi.org/10.1093/g3journal/jkac020

Caveats in application

We cannot assign clusters:

NSPHS is ≈1 cluster, i.e. Karesuando and Soppero

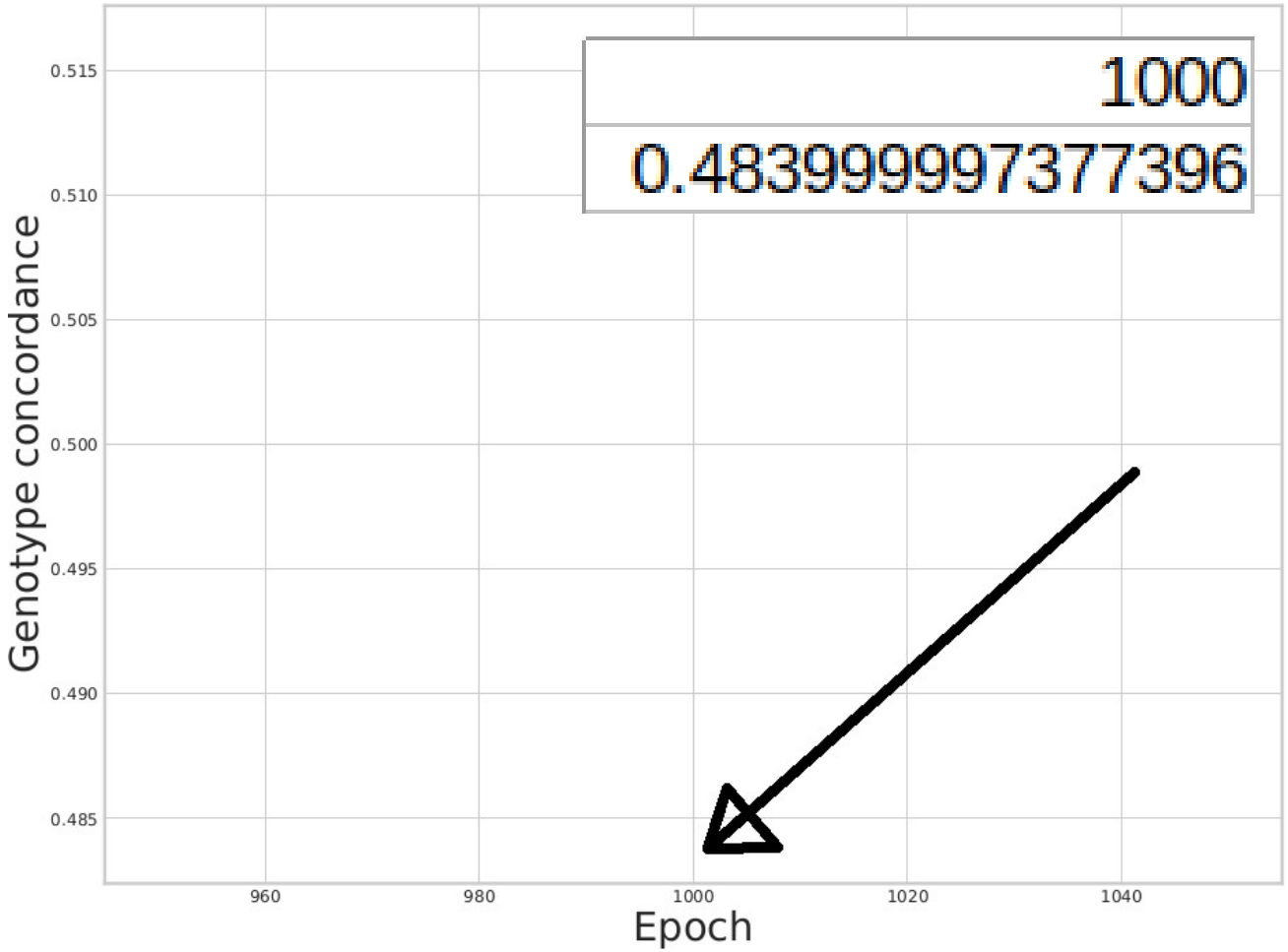
All we care about is that dimensionality reduction has reached equilibrium

Using 2 clusters anyway may allow us to do so

Maybe we only care about the trait predictions

https://en.wikipedia.org/wiki/Titanic_(1997_film)#/media/File:Titanic_(1997_film)_poster.png

Dimension reduction quality: other way



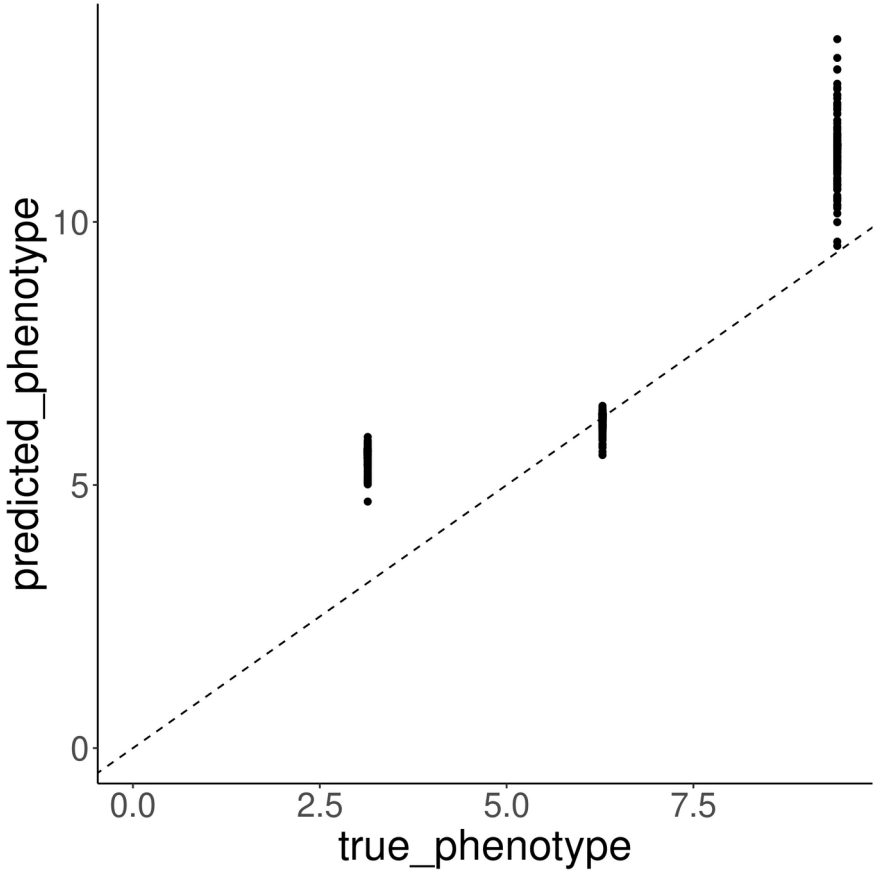
[ ] Plot the genotype concordance through time, #17

https://github.com/richelbilderbeek/gcaer/issues/17

Trait prediction score

How well are phenotypes predicted?

- … in time?
- … comparable between protein concentrations?



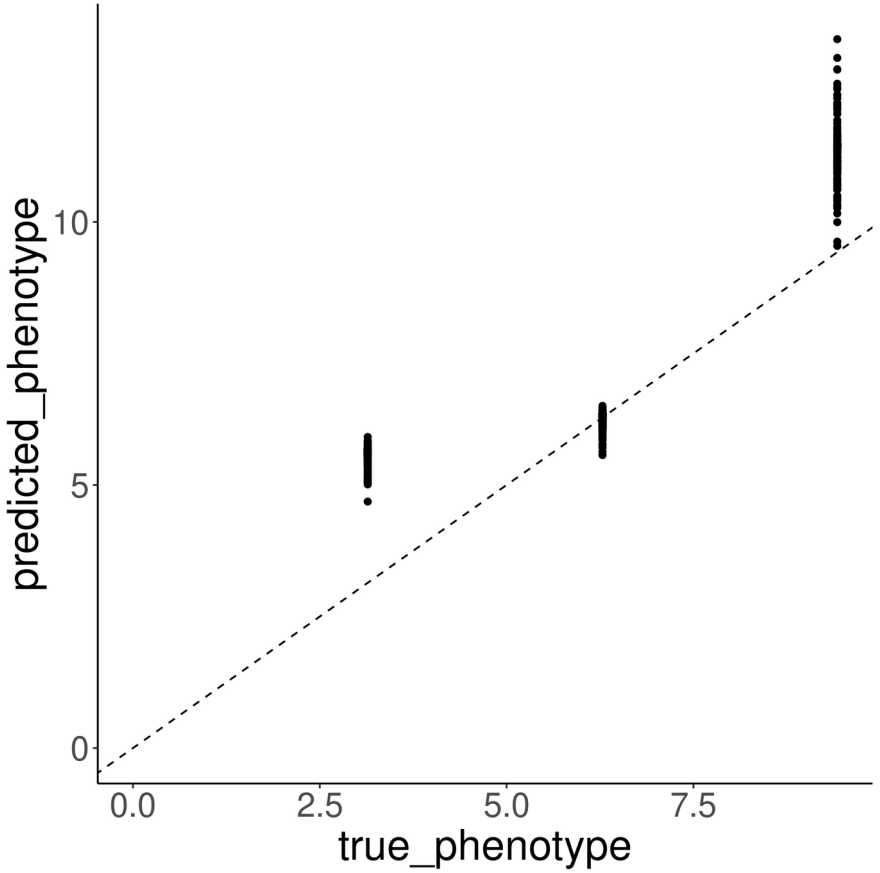
Trait prediction score 1/2: MSE

MSE: Mean squared error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean (vertical distance to identity line)²

Scale matters



Trait prediction score 2/2: NMSE

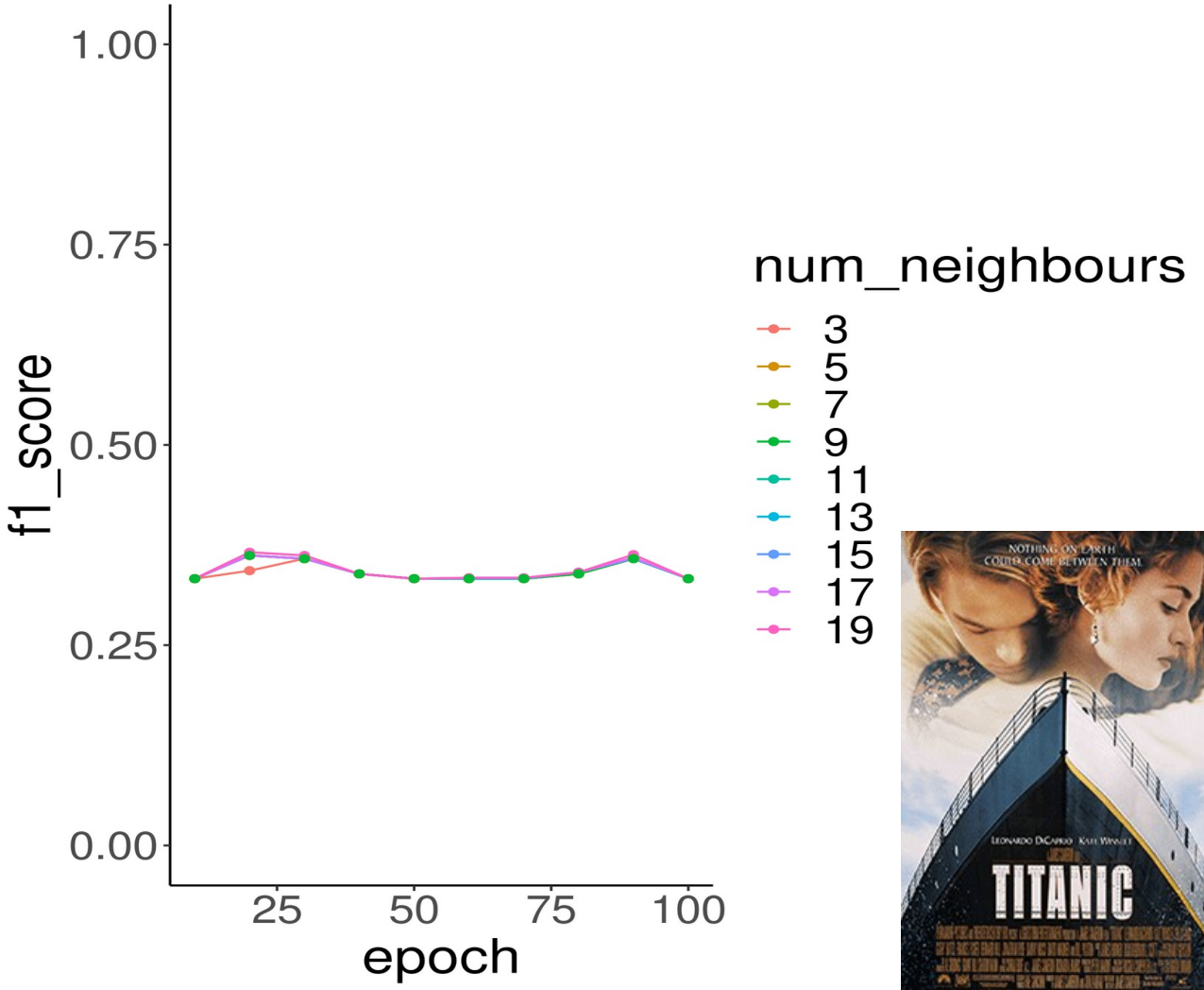
NMSE: Normalized Mean Squared Error

```
mean <- mean(true_values)
sd <- sd(true_values)
testthat::expect_true(sd > 0.0)
normalized_true_values <- (true_values - mean) / sd
normalized_estimated_values <- (estimated_values - mean) / sd
```

https://github.com/richelbilderbeek/gcaer/blob/master/R/normalize_true_and_estimated_values.R

Progress: get learning trajectory

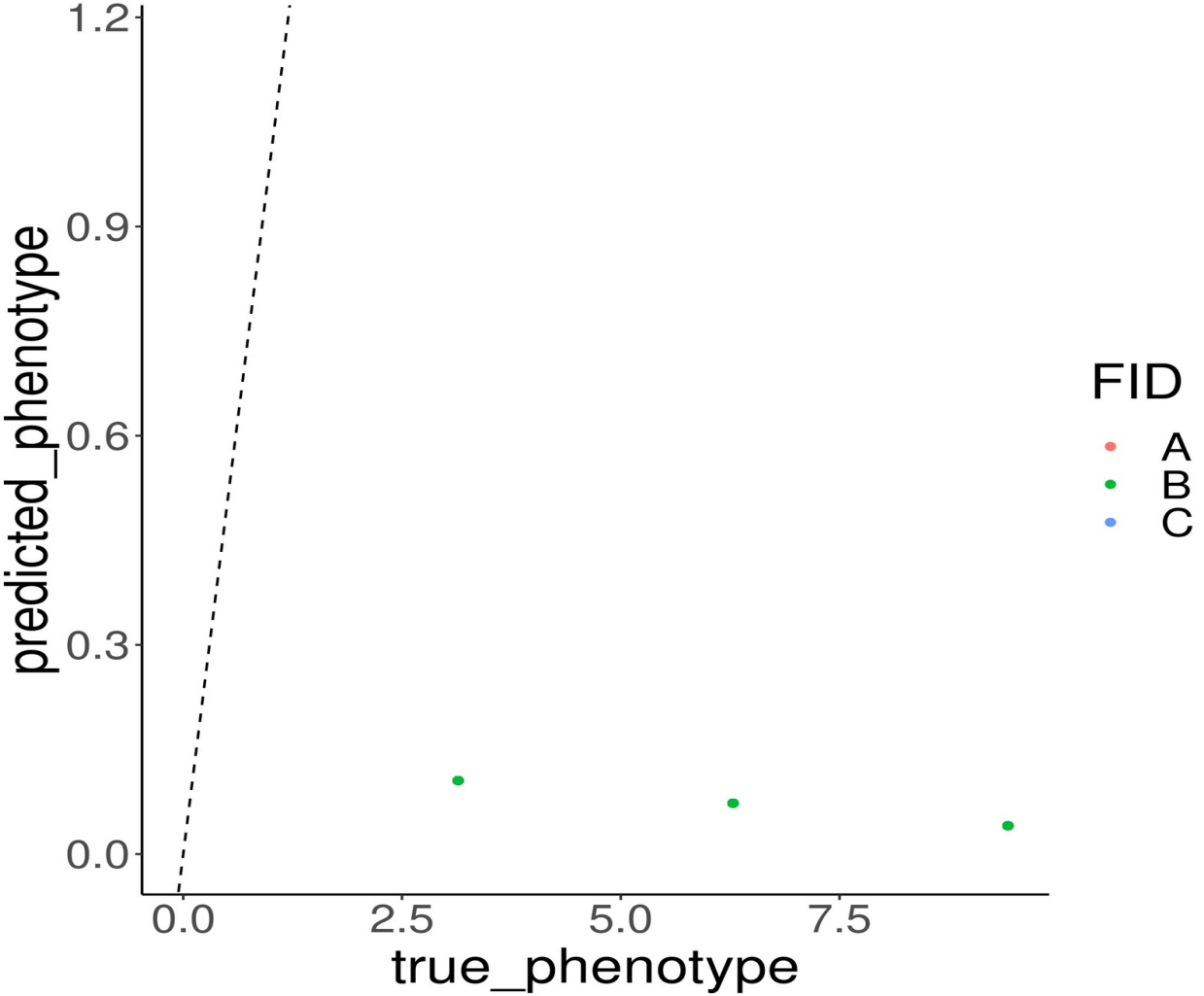
[x] Plot the F1 scores through time, #6



Progress: get learning trajectory

[x] Plot phenotype prediction through time, #21

[ ] Plot NMSE through time, #7

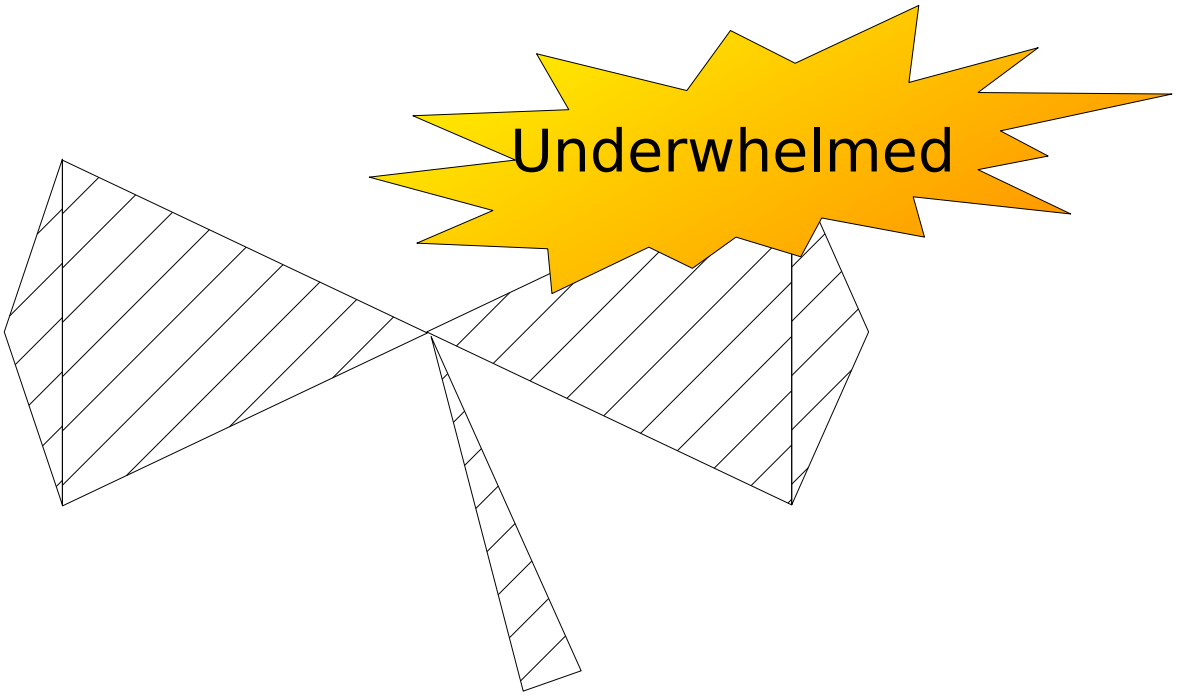


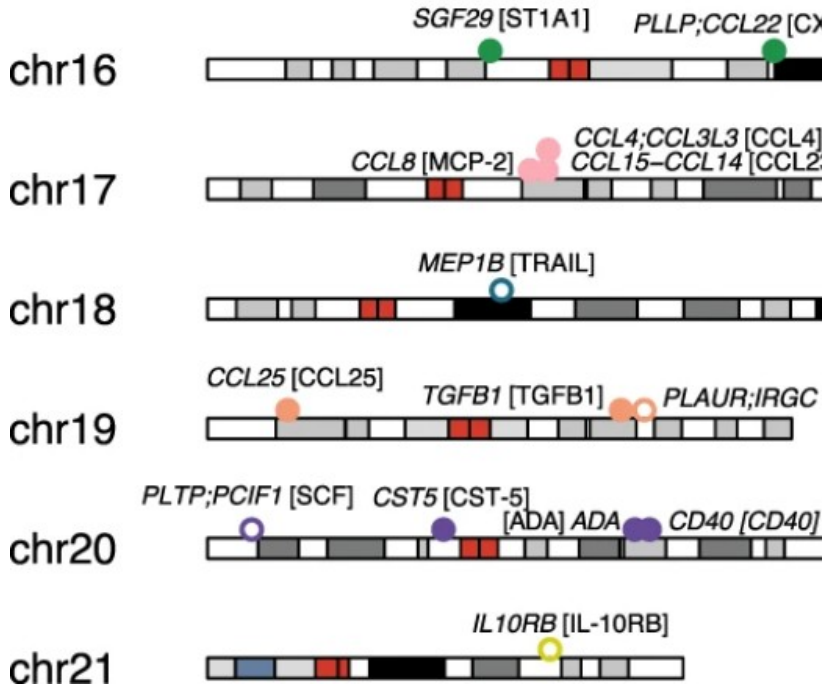
→ Show phenotype_prediction_through_time.gif

Conclusion

The autoencoder needs a hard problem

Hence, use NSPHS with useful genetic regions

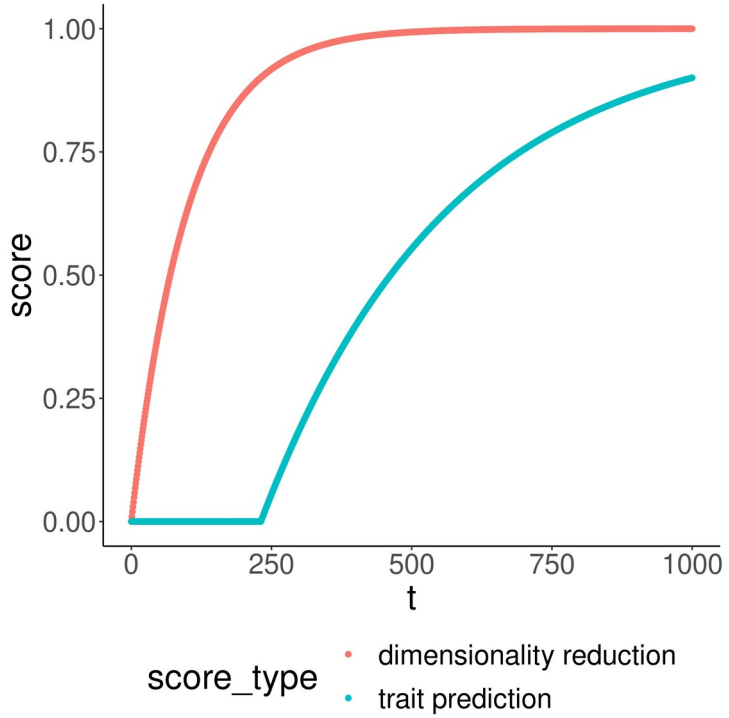




Conclusion

Need to determine when GCAE is done learning

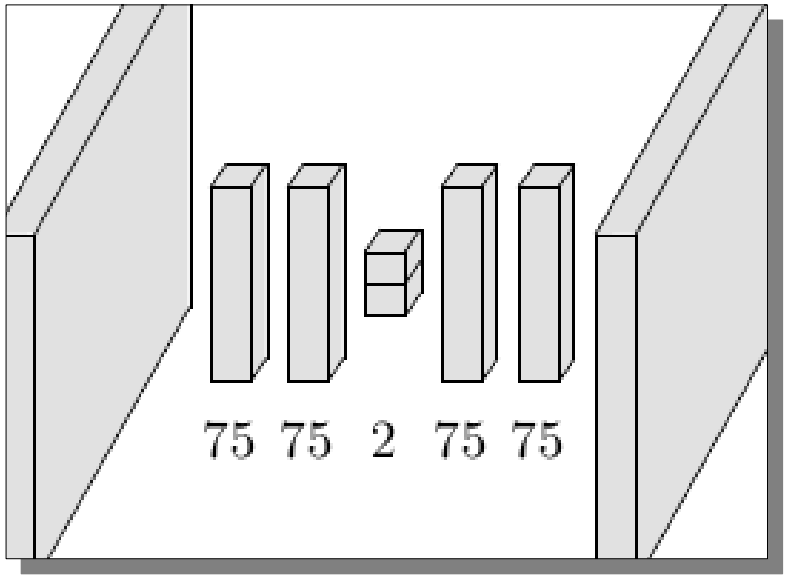
Hence, need to select/devise measures for that



Conclusion

The autoencoder then probably needs tuning

With a latent layer of 2 neurons (hence 2 dimensions), the autoencoder may underfit challenging data



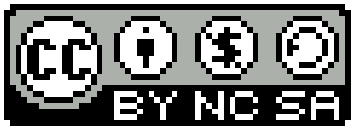
Discussion

We cannot predict where/where the autoencoder outperforms existing methods …

but it will be when:

- Data of sufficient complexity
- Non-linear relations between the principal components
- Common alleles
- (Noisy data)

Questions?



https://github.com/richelbilderbeek/science_presentation_20220305
https://youtu.be/ldwPcy263IU