CiBreed
Center for Integrated Breeding Research

Deep learning – an alternative for genomic prediction?

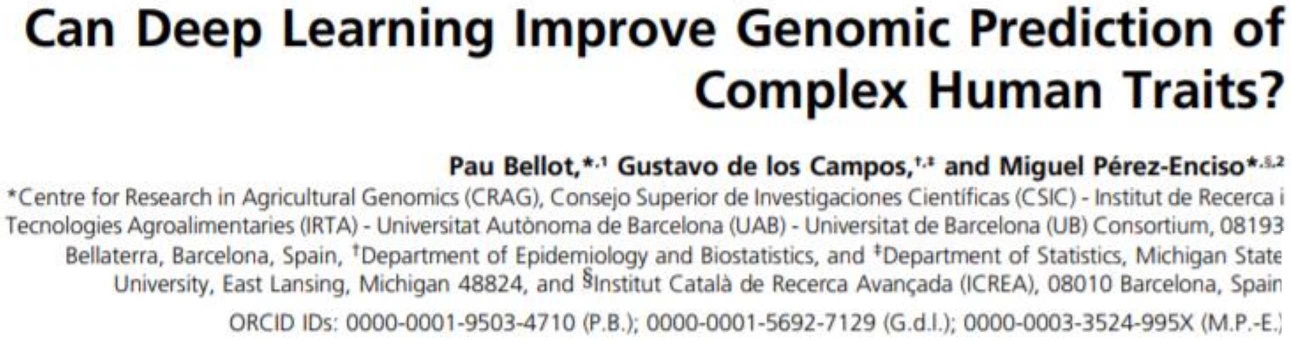
T. Pook[1], S. Herzog[2], J. Heise[3], H. Simianer[1]

[1] Department of Animal Sciences, Center for Integrated Breeding Research, University of Goettingen, 37075 Goettingen

[2] Max Planck Institute for Dynamics and Self-Organization, 37077 Goettingen

[3] IT Solutions for Animal Production (vit), 27283 Verden

Lots of people starting to use them!

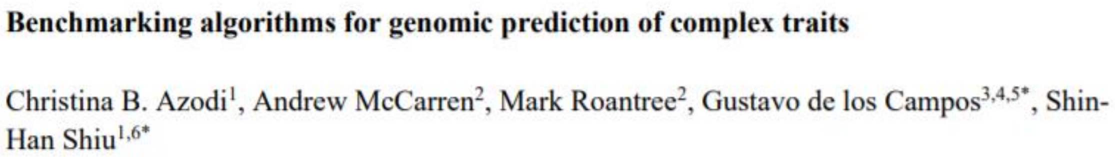


Can Deep Learning Improve Genomic Prediction of Complex Human Traits?

Pau Bellot,[*,1] Gustavo de los Campos,[†,‡] and Miguel Pérez-Enciso[*,§,2]

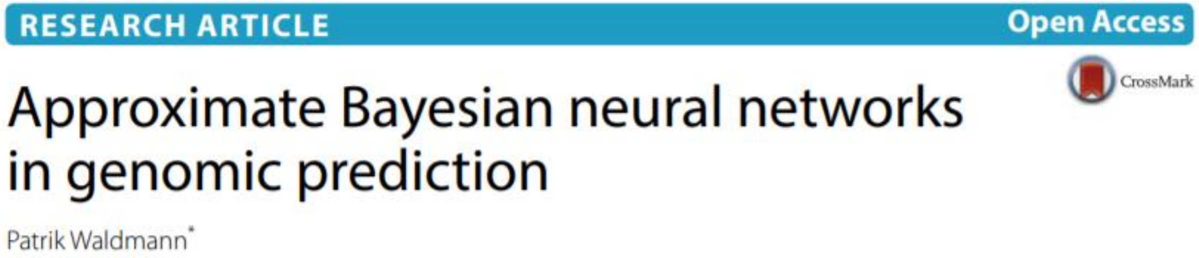
*Centre for Research in Agricultural Genomics (CRAG), Consejo Superior de Investigaciones Científicas (CSIC) - Institut de Recerca i Tecnologies Agroalimentaries (IRTA) - Universitat Autònoma de Barcelona (UAB) - Universitat de Barcelona (UB) Consortium, 08193 Bellaterra, Barcelona, Spain, †Department of Epidemiology and Biostatistics, and ‡Department of Statistics, Michigan State University, East Lansing, Michigan 48824, and §Institut Català de Recerca Avançada (ICREA), 08010 Barcelona, Spain

ORCID IDs: 0000-0001-9503-4710 (P.B.); 0000-0001-5692-7129 (G.d.l.); 0000-0003-3524-995X (M.P.-E.)



Benchmarking algorithms for genomic prediction of complex traits

Christina B. Azodi[1], Andrew McCarren[2], Mark Roantree[2], Gustavo de los Campos[3,4,5*], Shin-Han Shiu[1,6*]

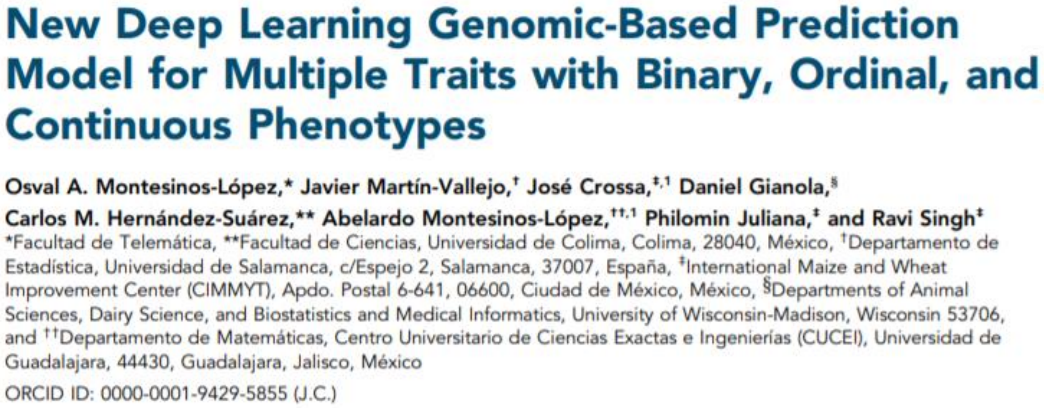


RESEARCH ARTICLE Open Access

CrossMark

Approximate Bayesian neural networks in genomic prediction

Patrik Waldmann[*]



New Deep Learning Genomic-Based Prediction Model for Multiple Traits with Binary, Ordinal, and Continuous Phenotypes

Osval A. Montesinos-López,[*] Javier Martín-Vallejo,[†] José Crossa,[‡,1] Daniel Gianola,[§] Carlos M. Hernández-Suárez,[**] Abelardo Montesinos-López,[††,1] Philomin Juliana,[‡] and Ravi Singh[‡]

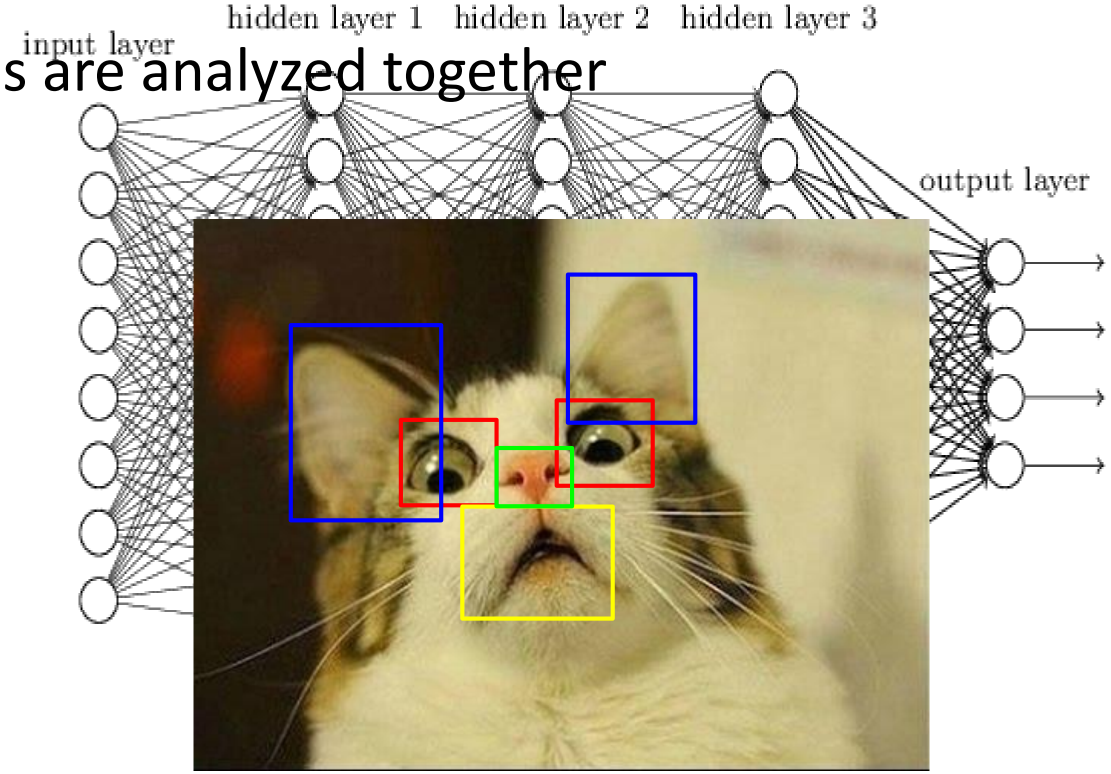
*Facultad de Telemática, **Facultad de Ciencias, Universidad de Colima, Colima, 28040, México, †Departamento de Estadística, Universidad de Salamanca, c/Espejo 2, Salamanca, 37007, España, ‡International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, Ciudad de México, México, §Departments of Animal Sciences, Dairy Science, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Wisconsin 53706, and ††Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, 44430, Guadalajara, Jalisco, México

ORCID ID: 0000-0001-9429-5855 (J.C.)

- Convolutional neural networks (CNN) do not work in this context!
- Other fields: CNN are the biggest reason for the rise of neural networks!

Neural networks are no black-box

- Fully-connected-layer
  - Nodes are connected to all nodes of the previous layer
- Convolutional-layer
  - Adjacent nodes are analyzed together

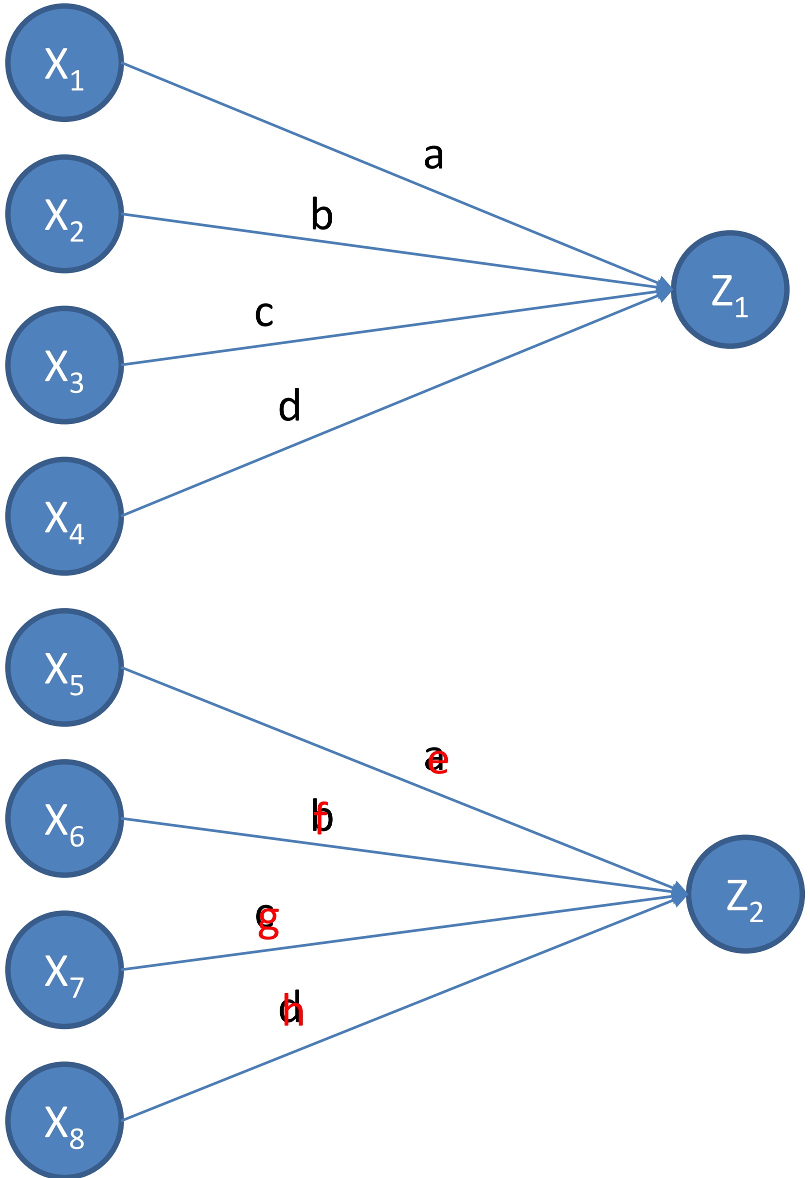


Problem with CNN

- Effects are assigned to specific sequences
  - e.g. for SNP-datasets 2201220
- BUT
  - Same sequences in different regions have different effects
  - Sequence is coding dependent (ancestral allele? / frequency based?)
  - What is between markers?

Our solution: Local convolutional layer

- Instead of using the same filter everywhere use local weightings
- For 50‘000 SNPs and 32 Nodes of a fully-connected-layer (FCL)
  - No CNN:
    - 1‘600‘000 parameters in the FCL
  - CNN (10 SNPs):
    - 10 parameters in the CNN
    - 160‘000 parameters in the FCL
  - Local CNN (10 SNPs):
    - 50‘000 parameters in the CNN
    - 160‘000 parameters in the FCL

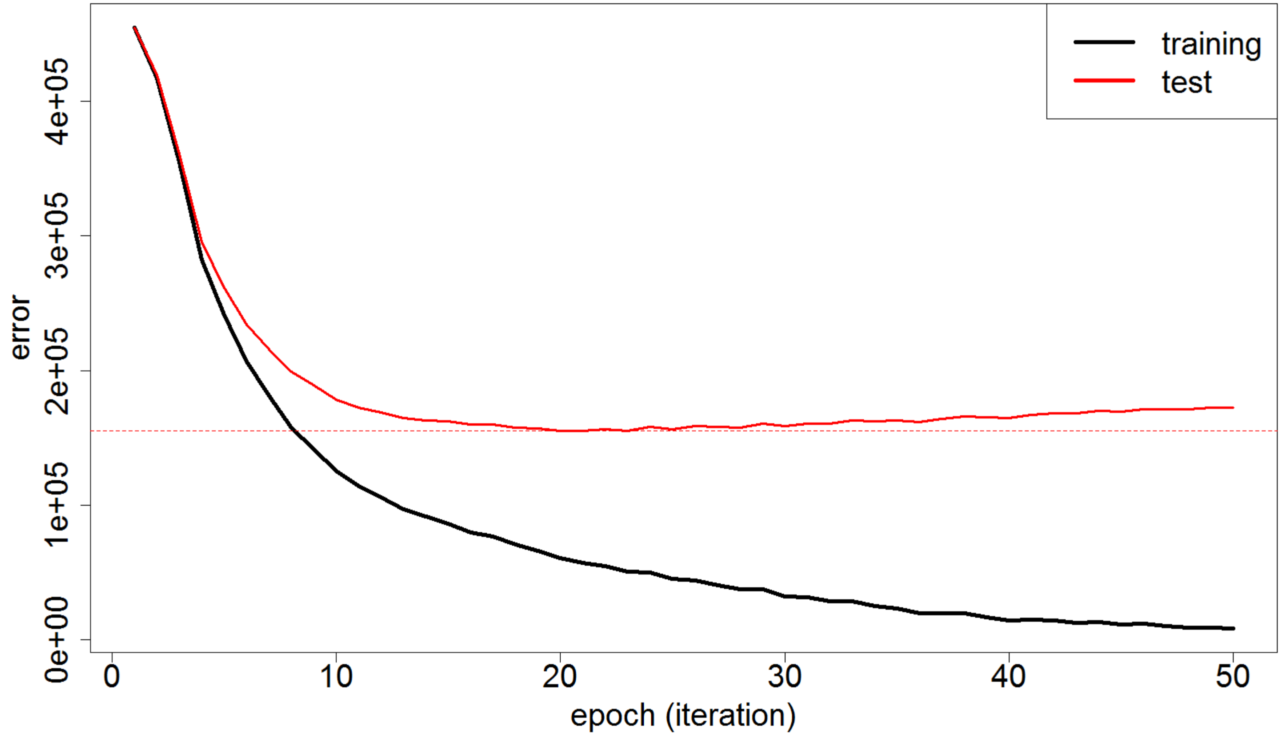


Data

- 10’501 bulls genotyped using a 50k chip
- Deregressed breeding values:
  - Milk yield (h^2 = 0.49)
  - Fat-kg (h^2 = 0.48)
  - Protein-kg (h^2 = 0.48)
  - Somatic cell score (h^2 = 0.23)
  - Non-Return-Rate (h^2 = 0.015)

Our model

- Local convolutional layer (15 SNPs, stride length = 10)
- Fully connect
- Fully connect



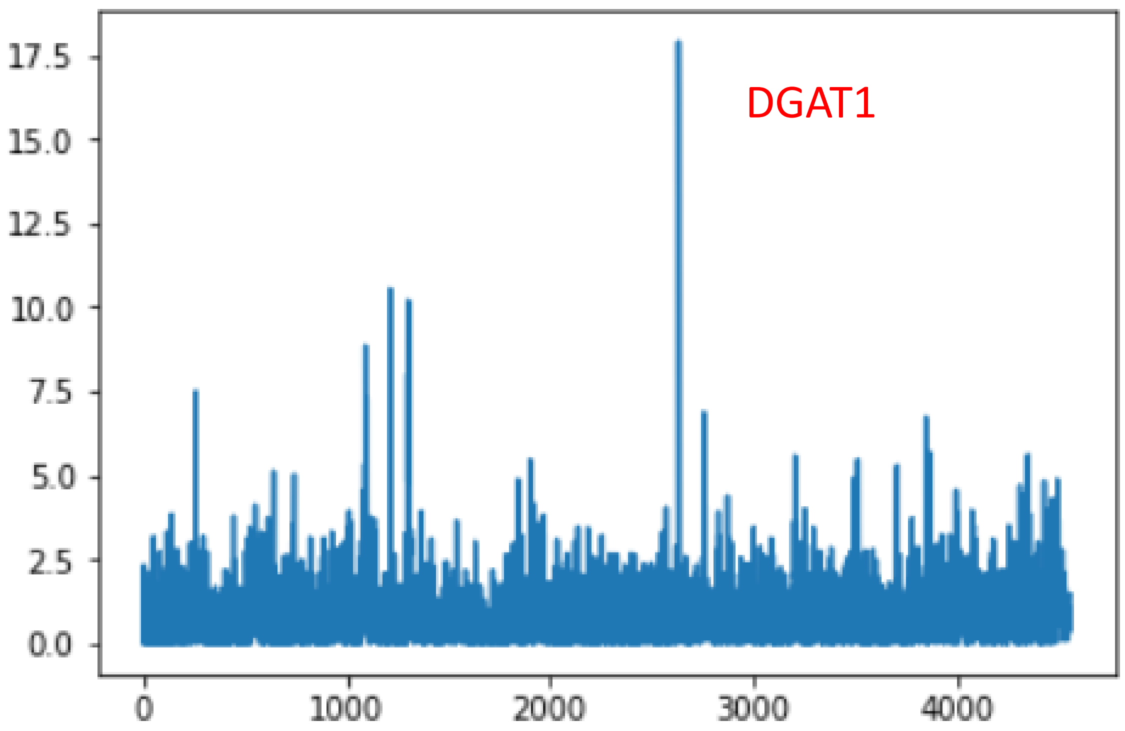
t

501)

- Use validatior n
  - How man
  - How many layers / nodes should be used
  - Target/optimization-function

Comparison to GBLUP

	GBLUP	Deep Learning	Change
Milk yield	0.830	0.834	+ 0.4 %
Fat-kg	0.809		
Protein-kg	0.822		
Somatic cell score	0.770		
Non-Return-Rate	0.658		



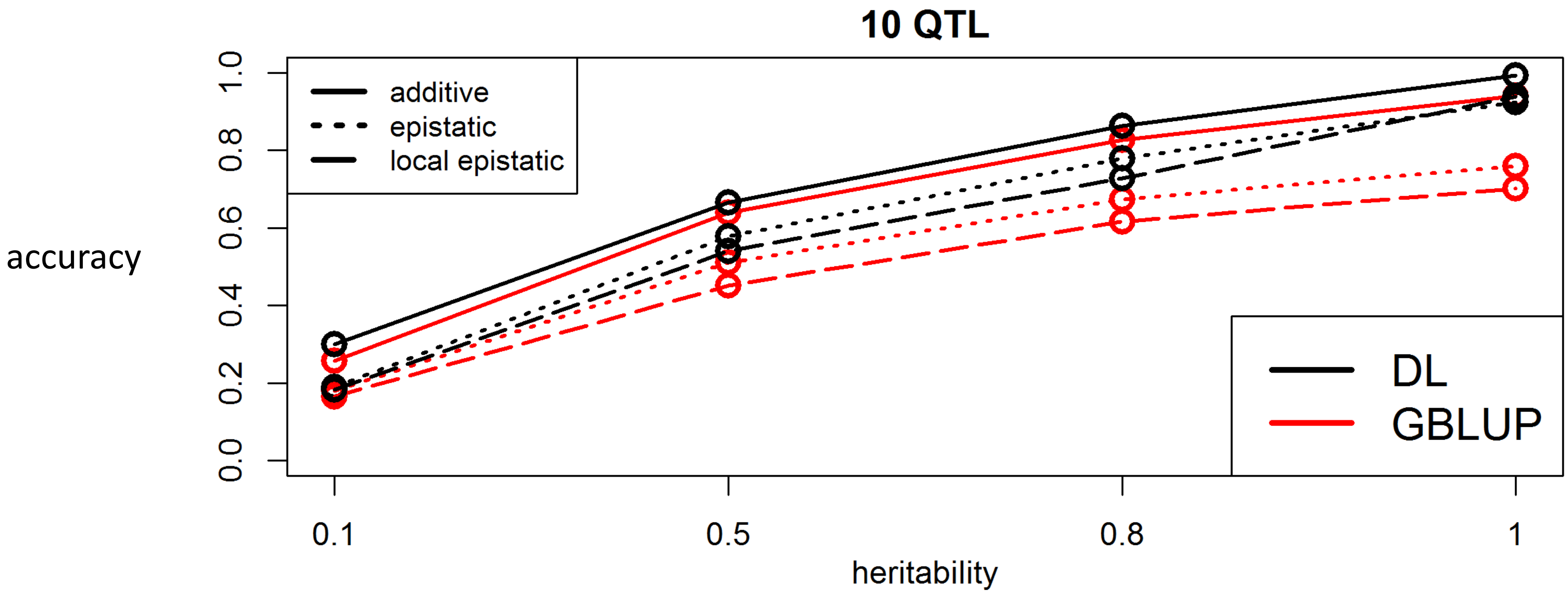
- Correlation of estimated breeding values a the test set
- DL is much worse for smaller training sets

Simulation study

- On what type of effect structures does Deep Learning work?
- Simulation of 10’000 animals
- 17 Traits of different complexity
  - 10 additive single marker QTL
  - 1’000 gamma distributed QTL caused by multiple physically linked QTL

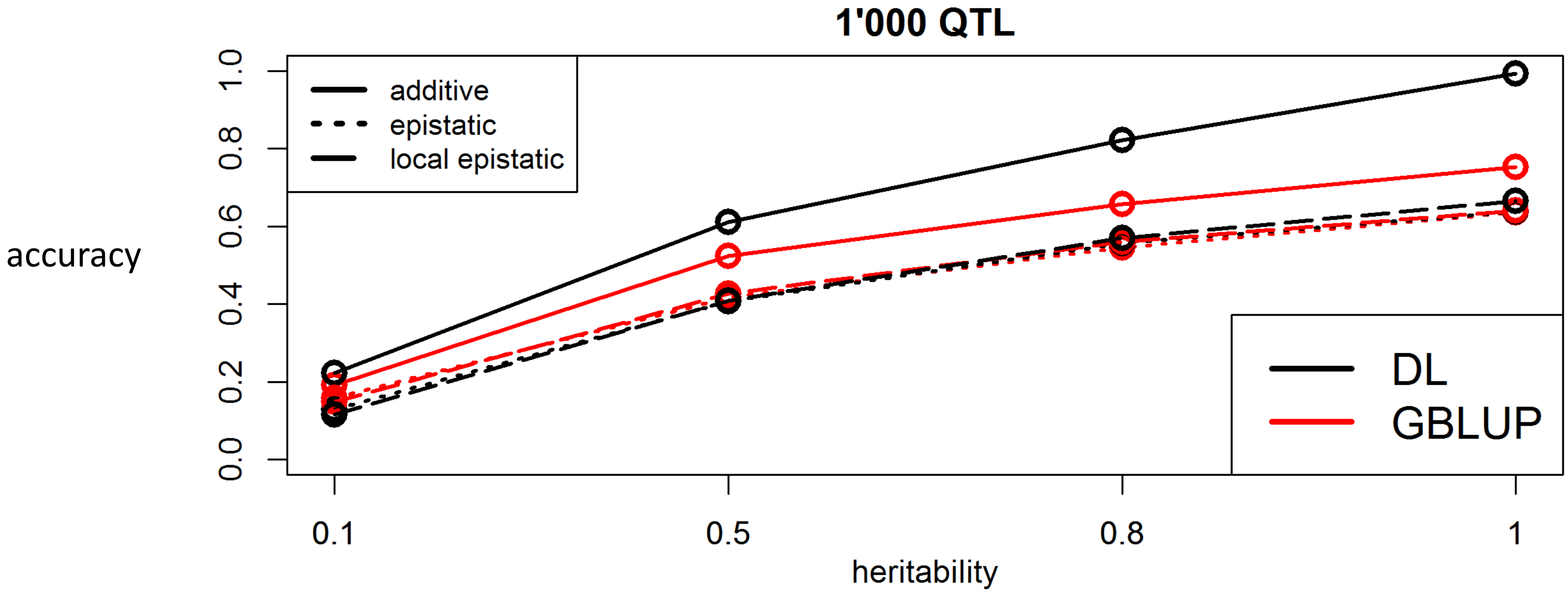
Low number of QTL

- Best performance for high heritability



High number of QTL

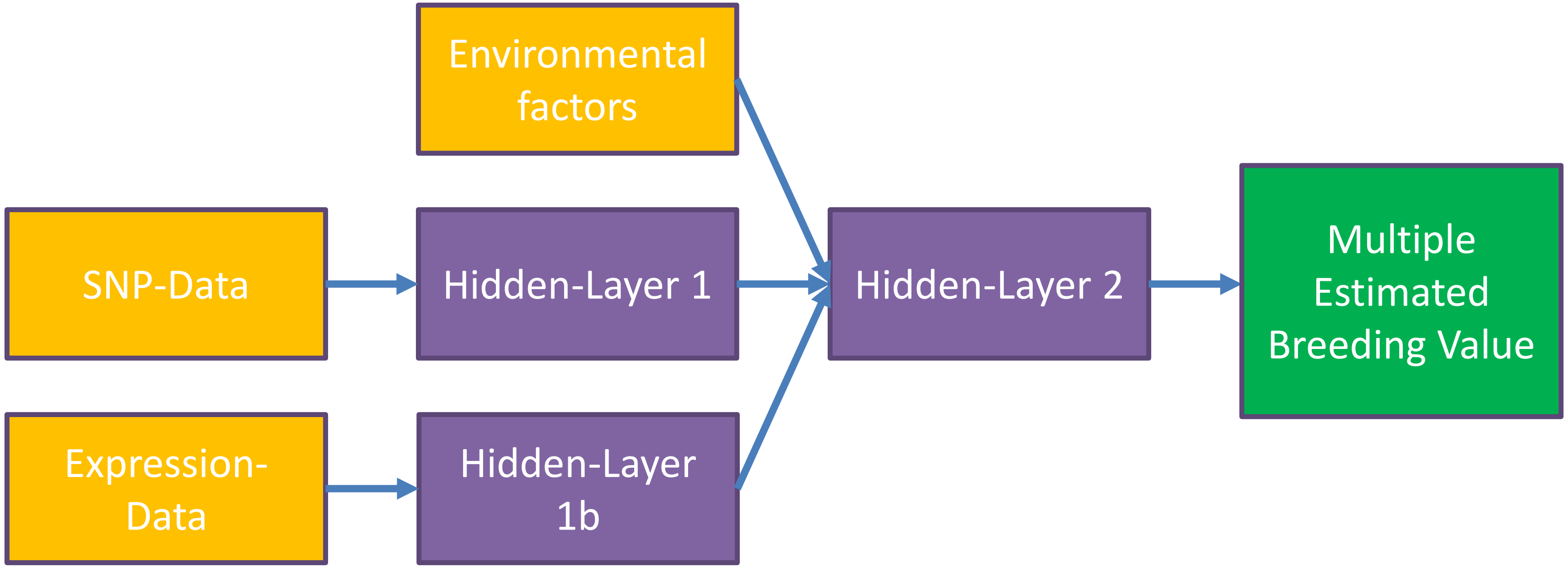
- Training set to small for highly complex traits?
- CNN do not excel in the local epistatic case



Further potential in genomic prediction

- Breeding values are additive by design!
- Genotypes of all individuals are needed!

- Phenotype prediction
- Expression data so far of limited usefulness
- High flexibility of input and output structure
- Linear scaling in computing time!



Acknowledgments

- MAZE: “Accessing the genomic and functional diversity of maize to improve quantitative traits”, BMBF Grant ID 031B0195
- FBF e.V. & vit

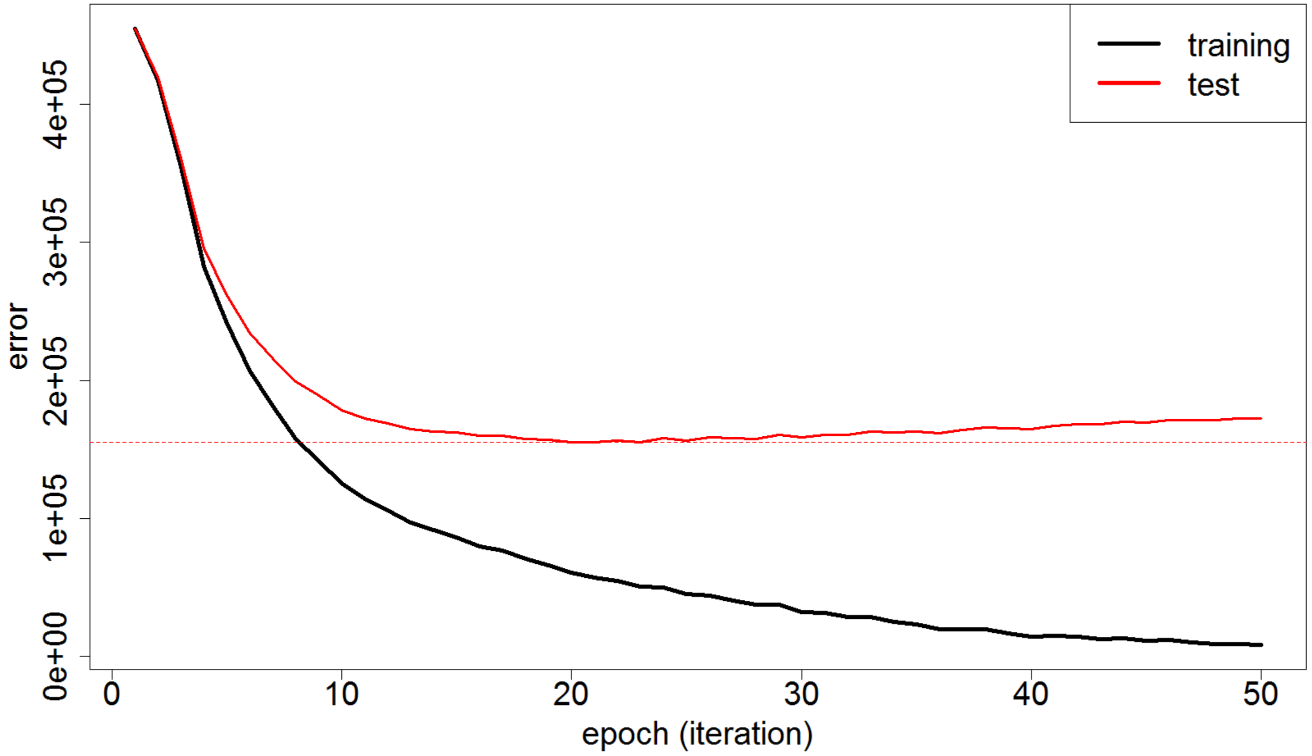
FBF Förderverein Bioökonomieforschung e.V.

vit Service & Daten aus einer Quelle

European Maize

Bundesministerium für Bildung und Forschung

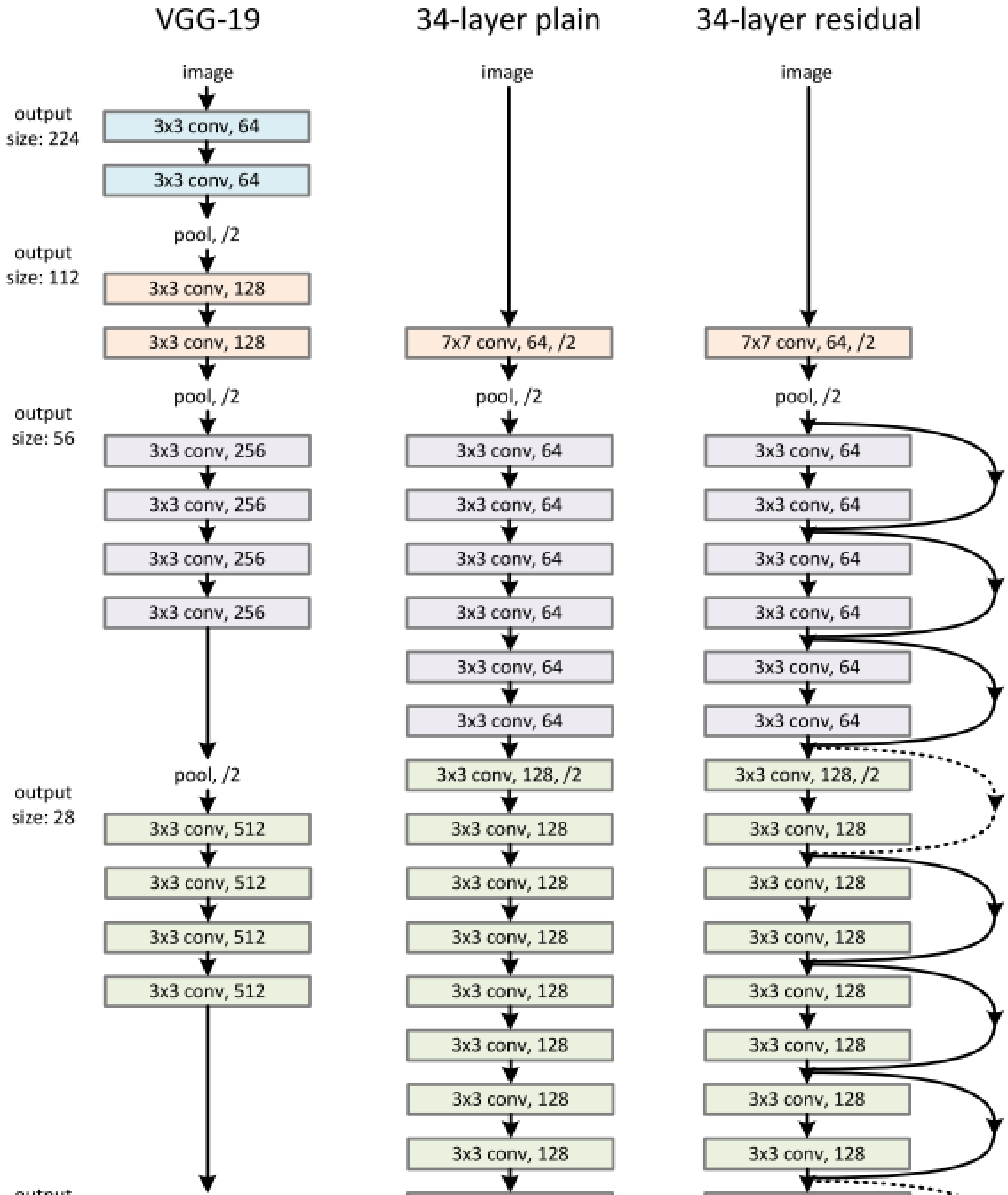
Results

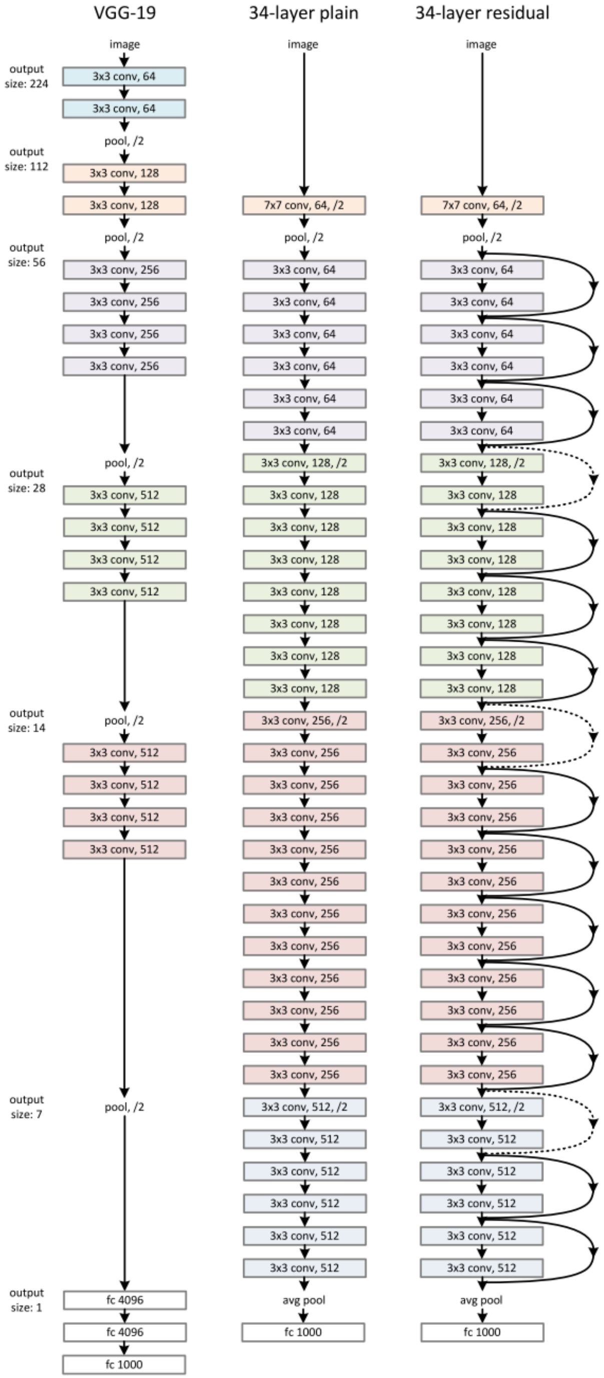


- Our current model contains 240.000 parameters
- Tendency of overfitting
- How to reduce overfitting or figure out when to stop

How to

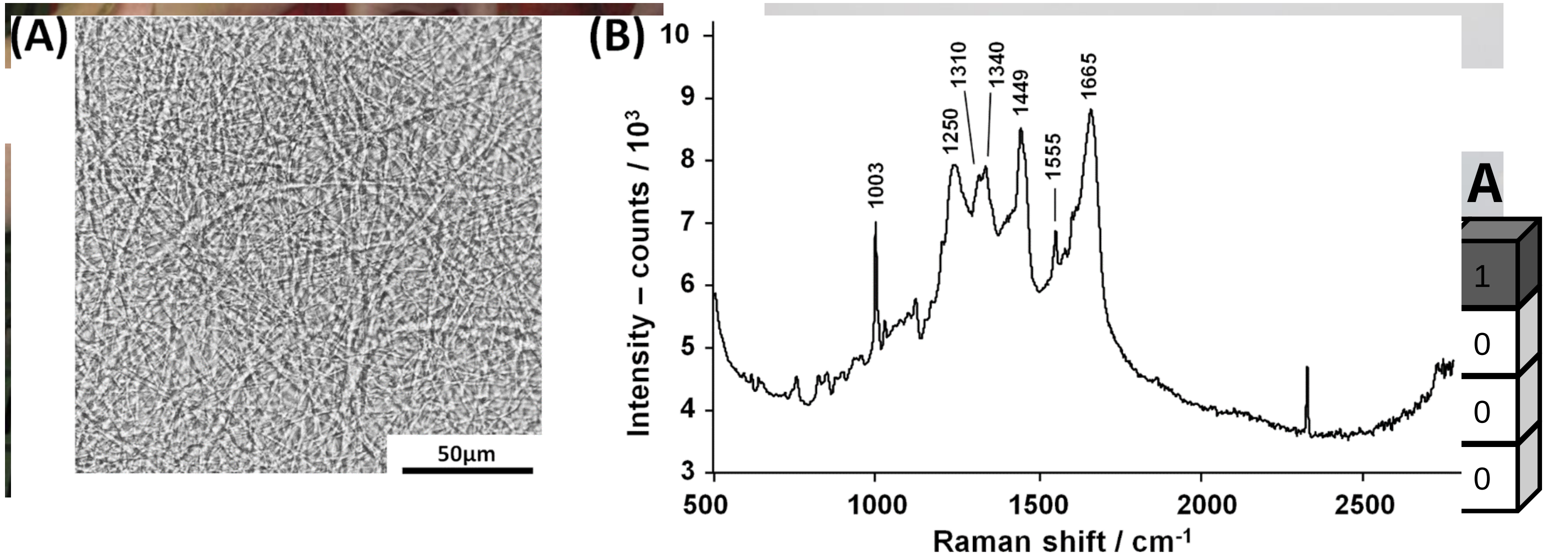
1. Build a
   - Inc at hand
2. Use ar lo some minor
   - Esp
3. Use ar
   - Crc
   - Ge
   - Esp data str




VGG-19
34-layer plain
34-layer residual
output size: 224
output size: 112
output size: 56
output size: 28
output size: 14
output size: 7
output size: 1
image
3x3 conv, 64
3x3 conv, 64
pool, /2
3x3 conv, 128
3x3 conv, 128
pool, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
fc 4096
fc 4096
fc 1000
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000

Application in genetics in general

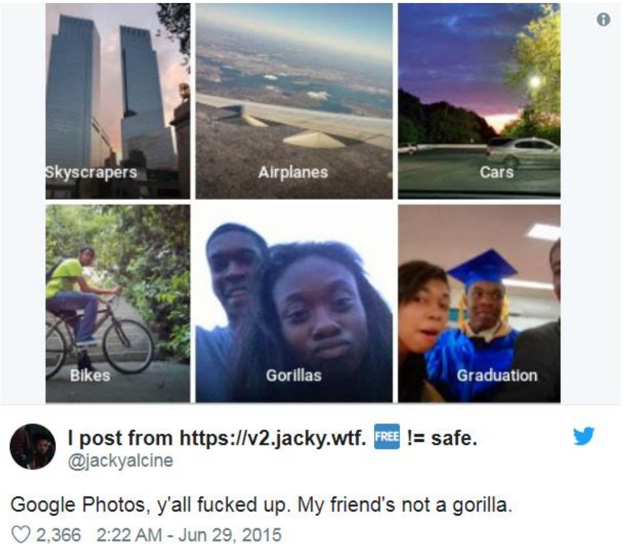
- Complex input and/or effect structure
- Spectral data (sexing of chicken, Galli et al. 2018)
- Phenotyping (Image and video analysis)
- Basically everything when working on sequence data
- Prediction of expression level (Washburn et al. 2019)



Expression levels?

Problem for genomic prediction

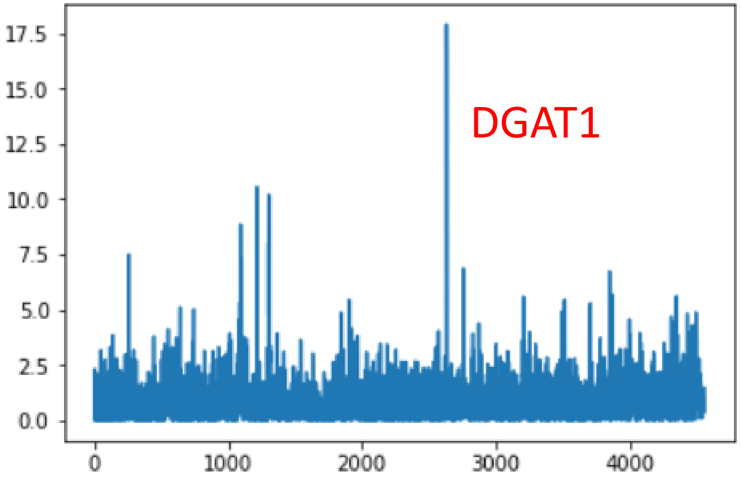
- No major gains when just using a SNP-dataset
- Old pipelines are already established
  - Model structure is far less understood // Black-Box
  - No reliabilities etc.
  - Goodness of fit outside of the training set
- Breeding for non-additive-effects in a random mating setting is not maximizing genetic gain



I post from https://v2.jacky.wtf. FREE != safe.
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

♡ 2,366 2:22 AM - Jun 29, 2015



Increasing the sample size

- Models are extremely data hungry:
- Generate additional data based on the already existing
- “Simple” way here:
  - Use same phenotype and some random mutations
  - Simulate a mating, use mean as phenotype
- Data augmentation
- Generative adversarial network:
  - Generate new data
  - Let the network determine which observation are simulated/real
  - Generate new data that would not be classified as fake in the previous model