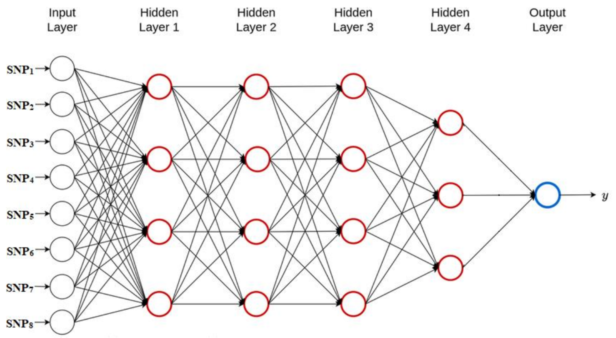


Optimizing Oil Palm Genomic Predictions with Artificial Neural Networks

David Cros, Lauriane Rouan, Daphné Navratil, Billy Tchounke,
Nicolas Leroy, Sandrine Le Squin, Najelaa Ulfah, Léifi Nodichao, Gregory Beurier

david.cros@cirad.fr

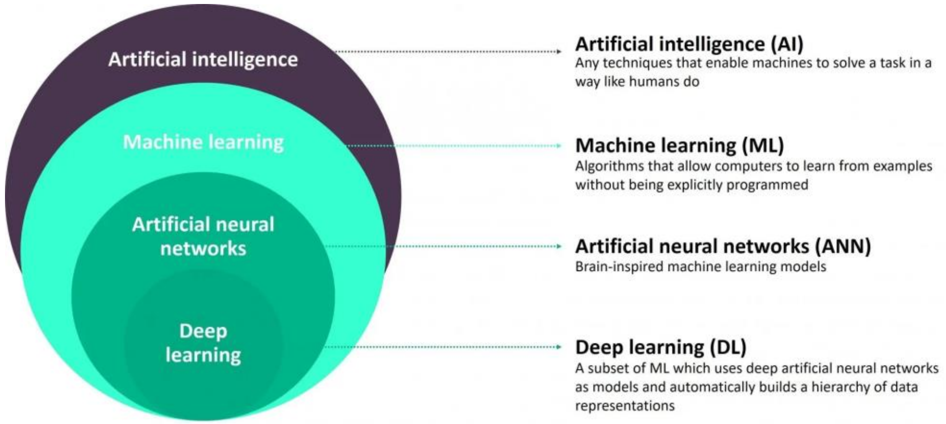
Computing facilities:

- High potential of genomic selection (GS) in perennial crops (long breeding cycles, low selection intensity)
- Promising results in oil palm, with r_{GS} = 0.25 – 0.75 depending on trait
- Still need to increase the accuracy of genomic predictions
- What about **innovative modeling approaches** ?

- High potential of genomic selection (GS) in perennial crops (long breeding cycles, low selection intensity)
- Promising results in oil palm, with r_{GS} = 0.25 – 0.75 depending on trait
- Still need to increase the accuracy of genomic predictions
- What about **innovative modeling approaches** ?

- Availability of large amount of heterogeneous data (phenotypes, high-throughput genotypes, NIRS, weather, …) = **machine learning** could be relevant
- Availability of computing resources = study and practical application of machine learning for GS feasible
- Promising results obtained for genomic predictions in various animal and plant species with machine learning, in particular artificial neural networks (ANN)

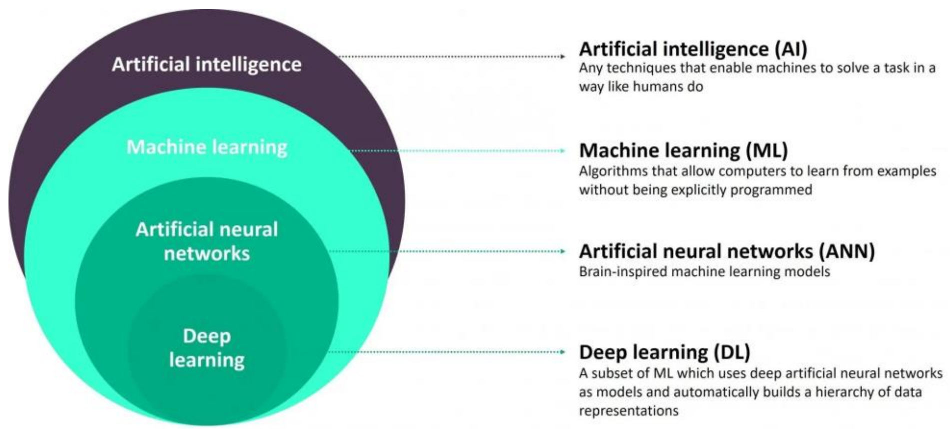
→ **Comparison of ANN and conventional statistical methods of genomic predictions**



- High potential of genomic selection (GS) in perennial crops (long breeding cycles, low selection intensity)
- Promising results in oil palm, with r_{GS} = 0.25 – 0.75 depending on trait
- Still need to increase the accuracy of genomic predictions
- What about **innovative modeling approaches** ?

- Availability of large amount of heterogeneous data (phenotypes, high-throughput genotypes, NIRS, weather, …) = **machine learning** could be relevant
- Availability of computing resources = study and practical application of machine learning for GS feasible
- Promising results obtained for genomic predictions in various animal and plant species with machine learning, in particular artificial neural networks (ANN)

→ **Comparison of ANN and conventional statistical methods of genomic predictions**



Optimal implementation of ANN can be challenging

→ **Study the effect of methodological aspects on ANN efficiency**

→ **Provide insights into how to achieve highest GS accuracies with ANN**

- **852 oil palm crosses** (69 717 individuals)
- complex dataset (structured in populations and families with varying size and levels of relatedness)
- phenotype: bunch production from 3 to 10 years old (FFB)
- genotype of cross parents and a sample of observed individuals for **22K SNP** (array)
- **2 experimental sites** in Indonesia

Site 1 (training)

Type de croisements par :		
Groupes génétiques	Populations	N
A x A	DELI x AN	1
A x B	AN x LM	10
A x B	DELI x (LM x YBI/SI)	23
A x B	DELI x LISOMBE KINSHASA	3
A x B	DELI x LM	243
A x B	DELI x NI	4
A x B	DELI x YBI	73
B x B	LM x NI	1
B x B	LM x YBI / SI_NI	1
B x B	NI	1
TOTAL		360

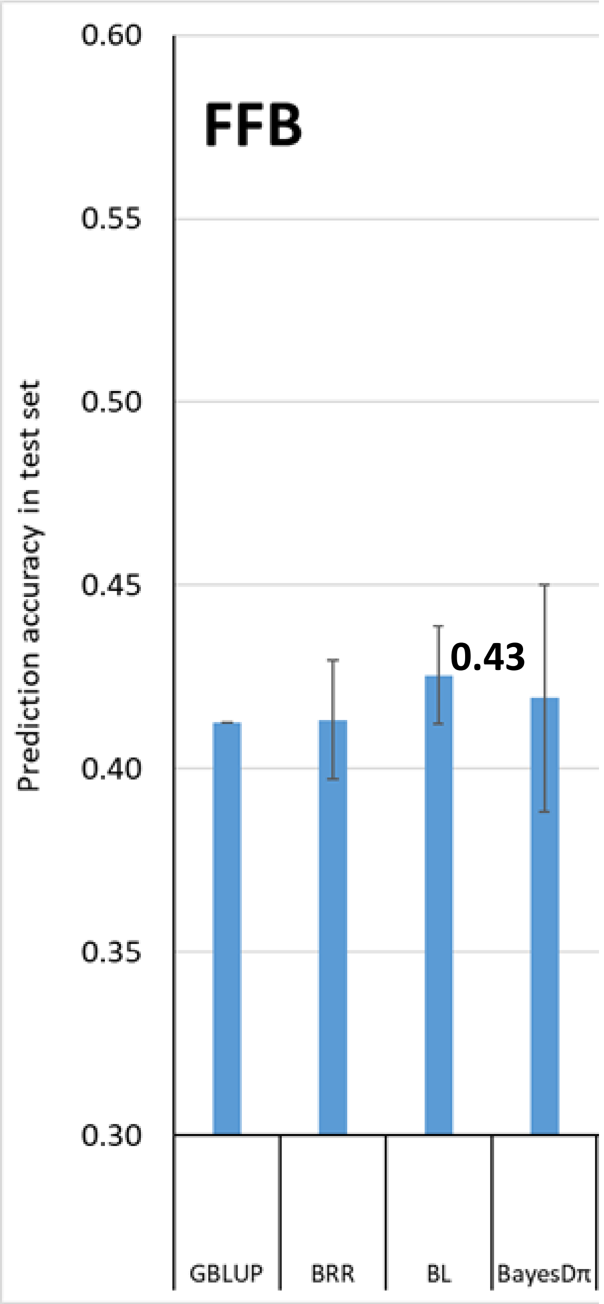
→ 688 training records

Site 2 (test)

Type de croisements par :		
Groupes génétiques	Populations	N
((AxB)xB) x (AxB)	(DELIxLM)xNI_x_DELIxNI	1
((AxB)xB) x (AxB)	(DELIxLM)xYBI_x_DELIxNI	3
((AxB)xB) x A	(DELIxLM)xLM_x_DELI	14
((AxB)xB) x A	(DELIxLM)xNI_x_DELI	1
((AxB)xB) x A	(DELIxLM)xYBI_x_DELI	3
((AxB)xB) x B	(DELIxLM)xNI_x_NI	3
((AxB)xB) x B	(DELIxLM)xYBI_x_NI	1
(AxB) x (AxB)	DELIxNI_x_DELIxYBI?	1
(AxB) x B	ANxNI_x_LM	3
(AxB) x B	ANxNI_x_YBI	2
(AxB) x B	DELIxNI_x_LISOMBE KINSHASAxLM	1
(AxB) x B	DELIxNI_x_LISOMBE KINSHASA	1
(AxB) x B	DELIxNI_x_LM	15
(AxB) x B	DELIxNI_x_LMxYBI/SI	5
(AxB) x B	DELIxNI_x_NIxLM	13
(AxB) x B	DELIxNI_x_YBI	6
A x (AxB)	DELI_x_DELIxYBI?	5
A x B	ANxDELI_x_LM	31
A x B	ANxDELI_x_YBI	21
A x B	DELI_x_LISOMBEKINSHASA	6
A x B	DELI_x_LISOMBEKINSHASAxLM	10
A x B	DELI_x_LM	188
A x B	DELI_x_LMxYBI/SI	21
A x B	DELI_x_NIxLM	15
A x B	DELI_x_NIxYBI	14
A x B	DELI_x_YBI	86
B x B	LM_x_NI	4
B x B	LM_x_YBI	4
B x B	LMxYBI/SI_x_NI	3
B x B	NI_x_NIxLM	11
TOTAL		492

→ 492 test records

Prediction accuracy of conventional methods in test set:



…the base artificial neural network: the **multi-layer perceptron (MLP)**

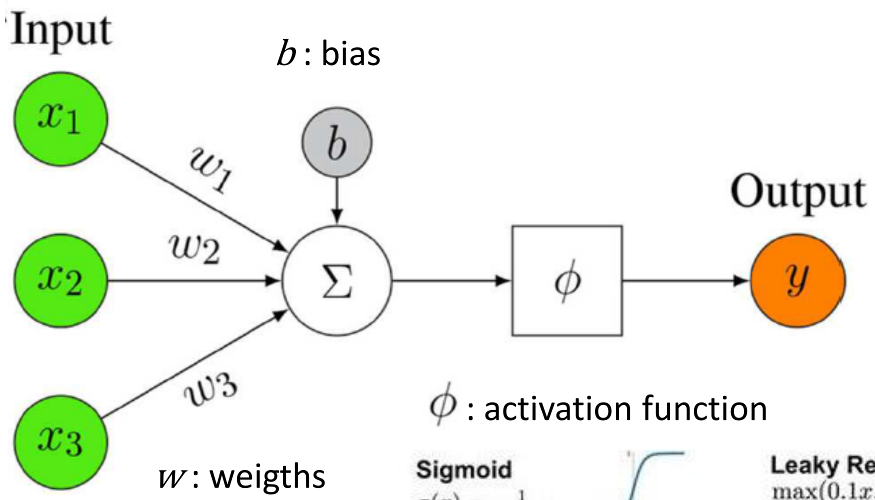
Single neuron:

b : bias

SNP1

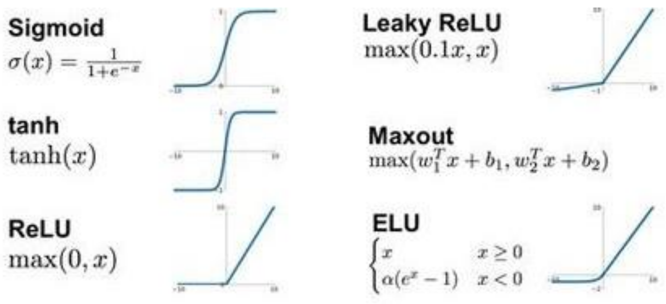
SNP2

SNP3



w : weigths

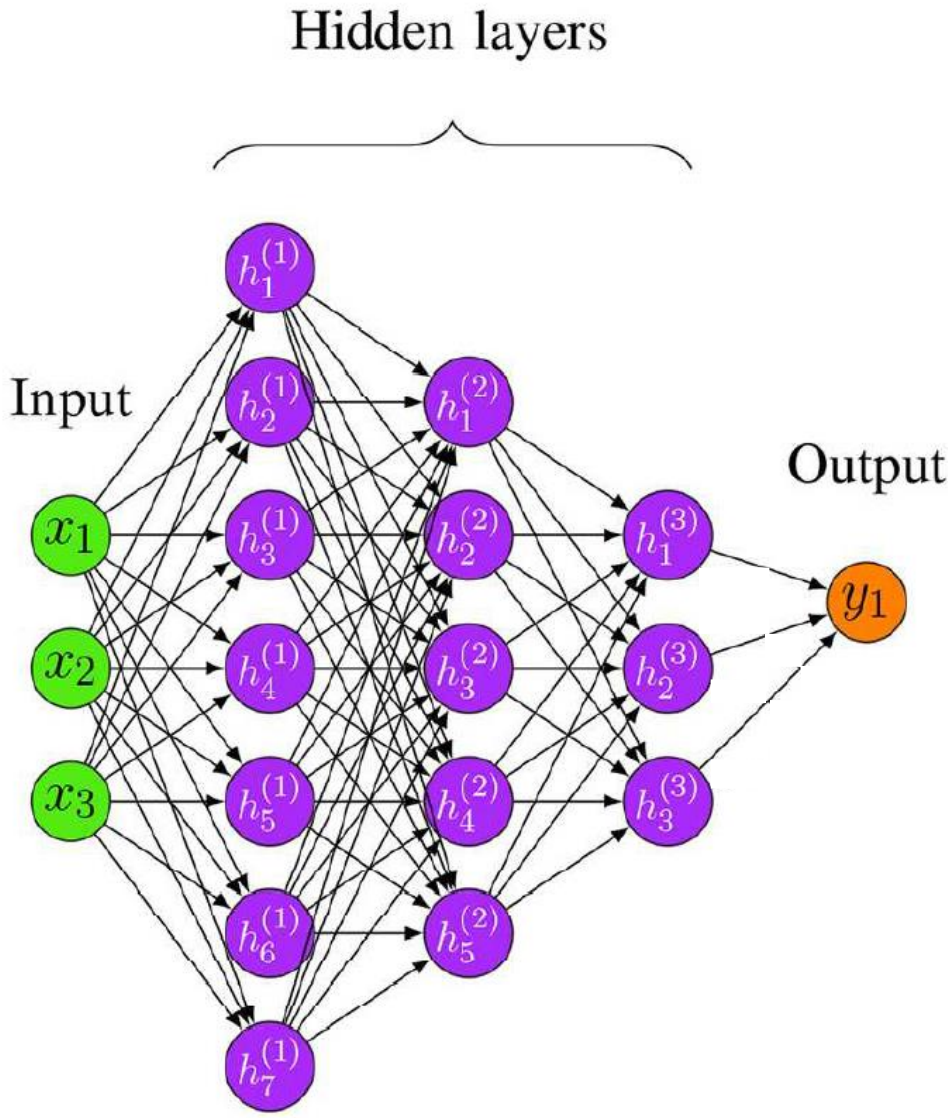
ϕ : activation function



etc.

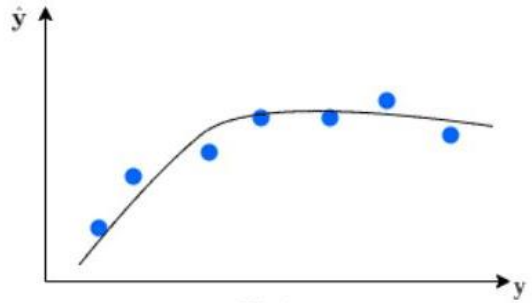
$$\rightarrow \quad y = \phi\left(\sum_{i=1}^{n} x_i w_i + b\right)$$

Network:

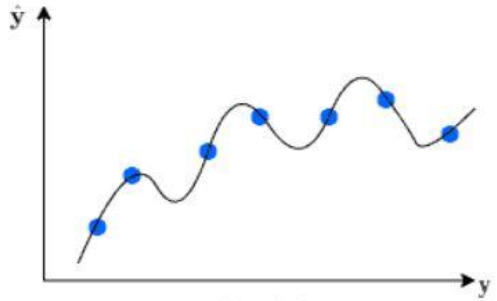


Emmer-Streib et al 2020

Prevention of overfitting:

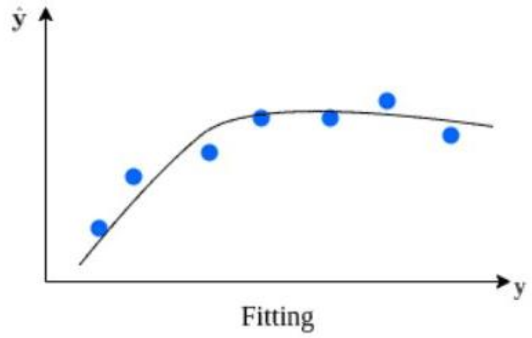


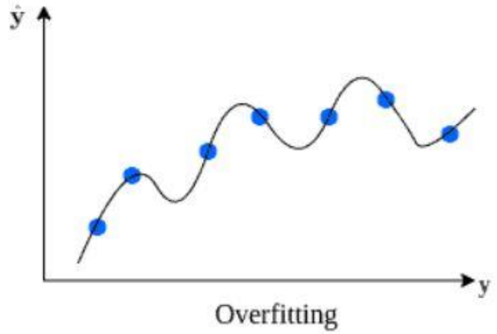
Fitting

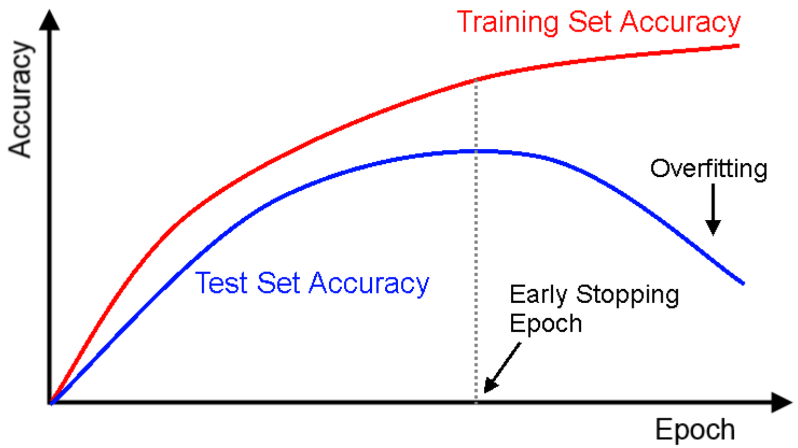


Overfitting

Prevention of overfitting:

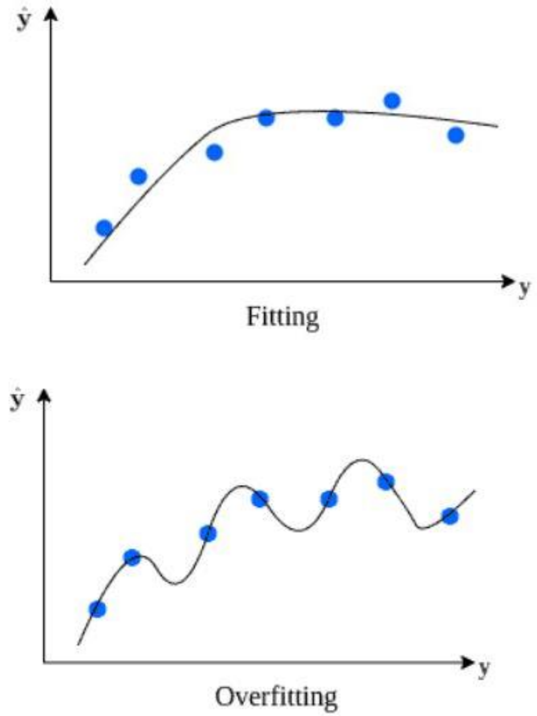


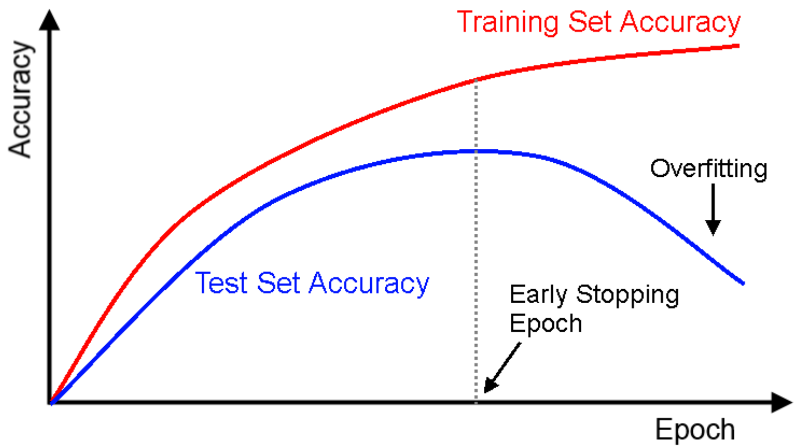




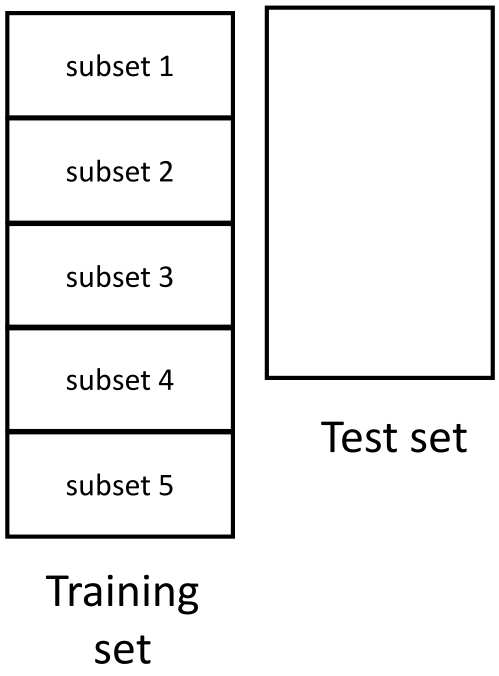
- **Divide training data into training and validation subsets** and use loss value in validation subset to identify optimal epoch (**early-stopping**)

Prevention of overfitting:

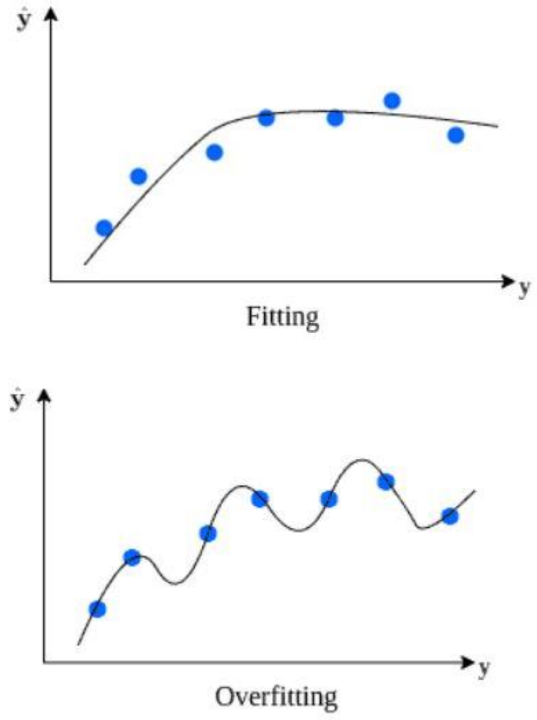


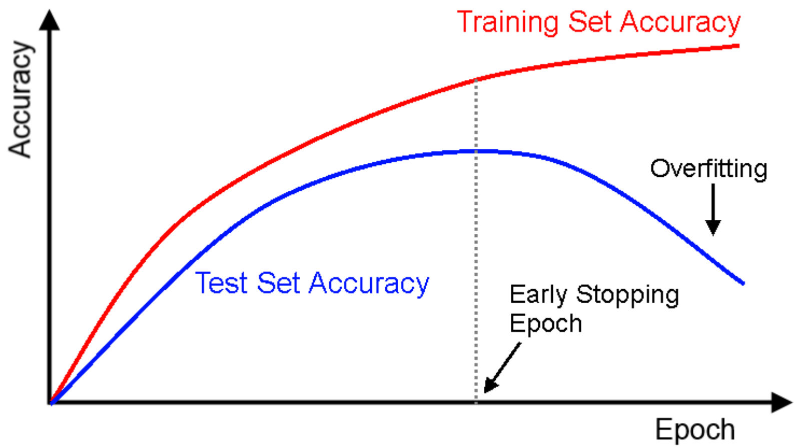


- **Divide training data into training and validation subsets** and use loss value in validation subset to identify optimal epoch (**early-stopping**)

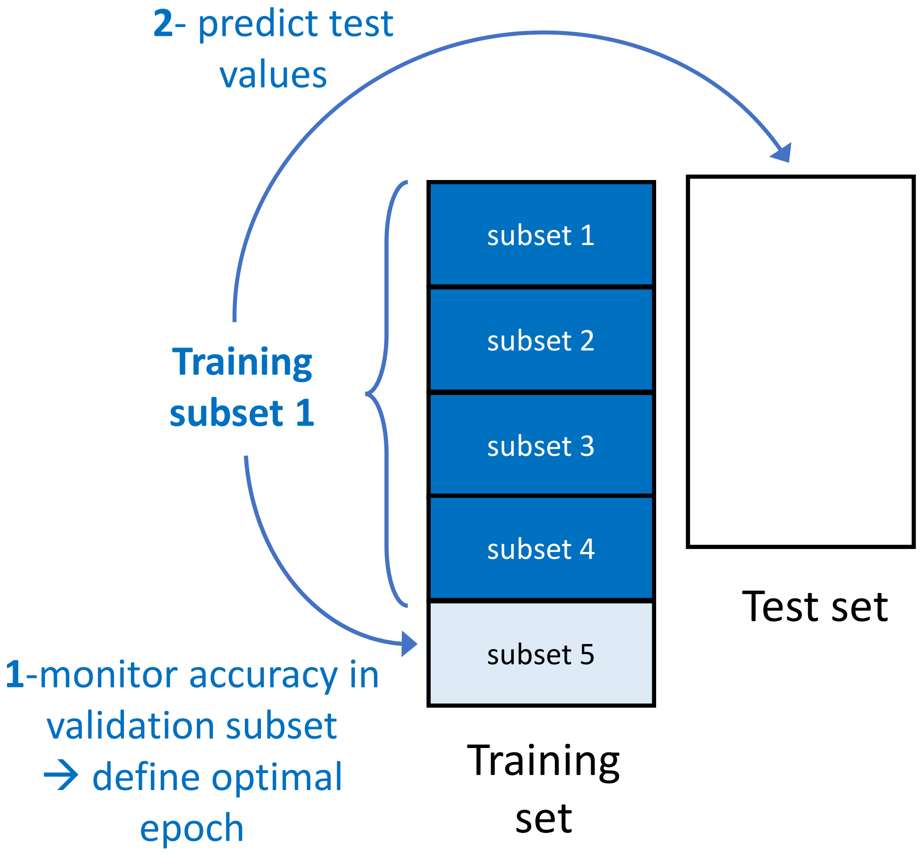


Prevention of overfitting:



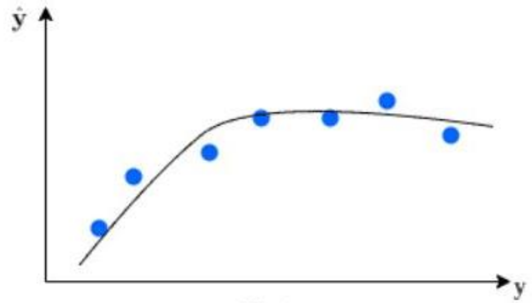


- **Divide training data into training and validation subsets** and use loss value in validation subset to identify optimal epoch (**early-stopping**)

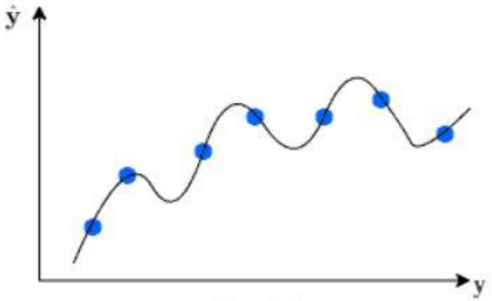


3- repeat steps 1 and 2 with the four other validation subsets

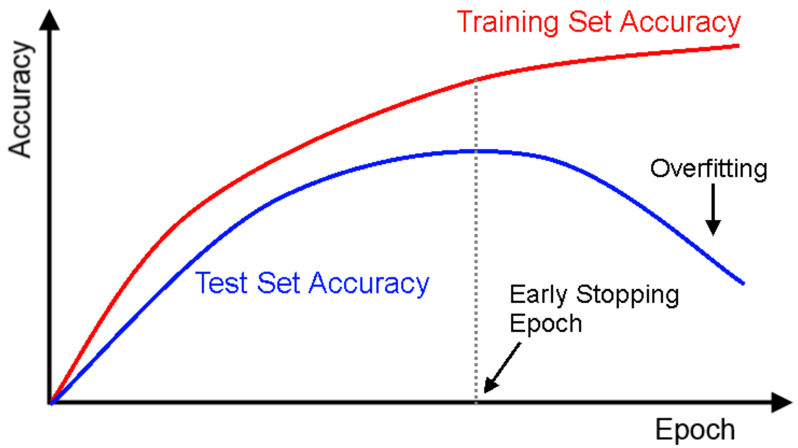
Prevention of overfitting:



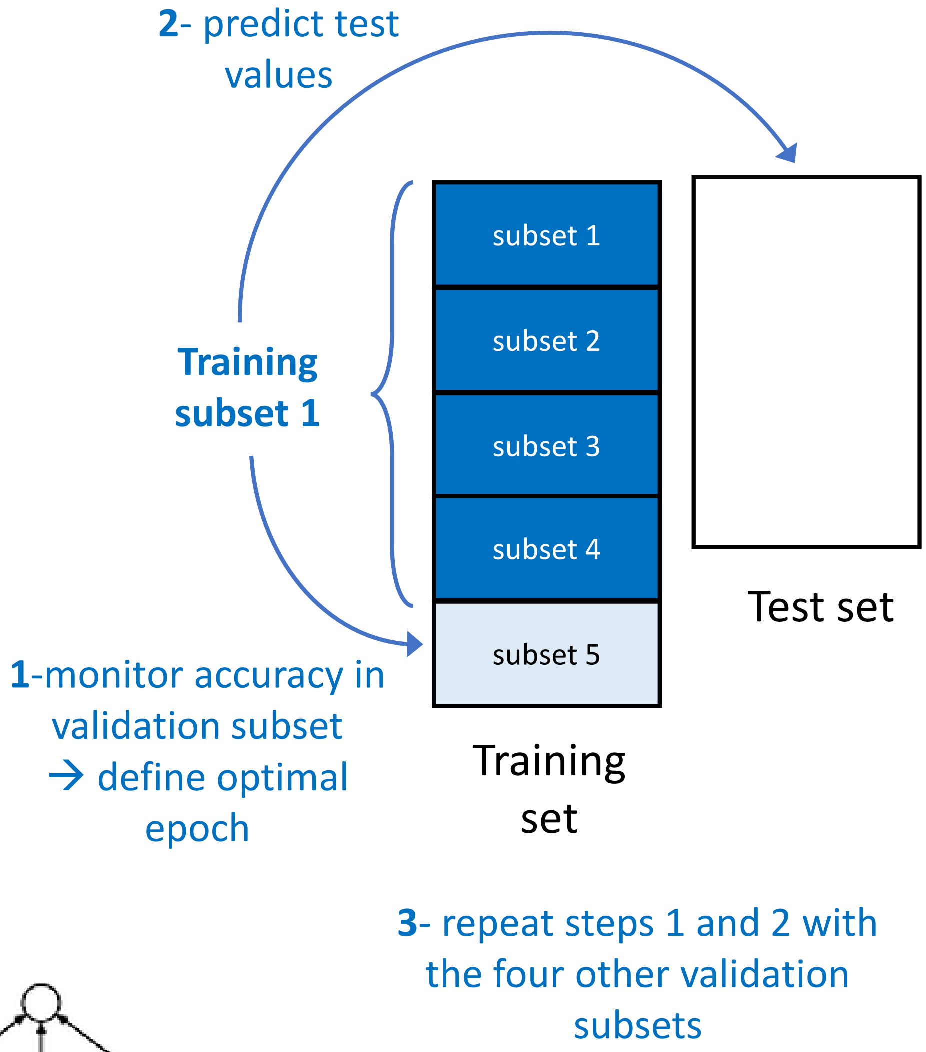
Fitting



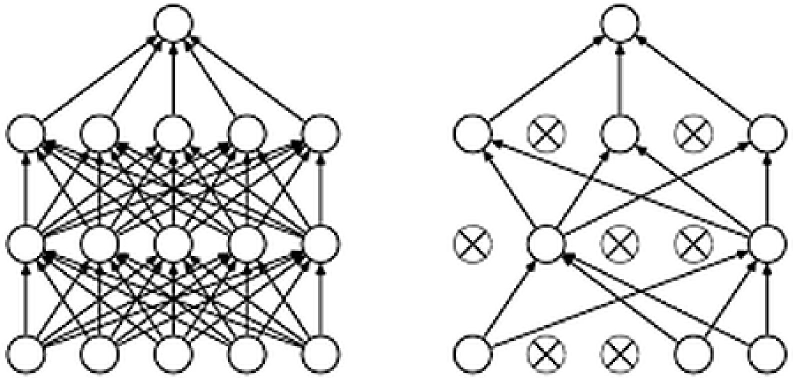
Overfitting



- **Divide training data into training and validation subsets** and use loss value in validation subset to identify optimal epoch (**early-stopping**)



- **Use regularization techniques – example: dropout**



(a) Standard Neural Net (b) After applying dropout.

Many possible MLP models:
- **architecture** (number of layers, number of neurons per layer)
- **hyper-parameters** (learning rate, regularization parameters [dropout, l1, l2], activation function, etc.)

Question 1. What is the variability in r_{test} among MLP ?

Predictions made for each training/validation subsets

Question 2. How to compute r_{test}: $\overline{cor(y,\hat{y})}$ ou $cor(y,\bar{\hat{y}})$?

Initial weights and biases generally fixed randomly
Dropout (random sampling of neurons to switch off)
Random definition of batches

→ **ANN non-deterministic methods**

Question 3. What is the variability in r_{test} for a given MLP and dataset ?

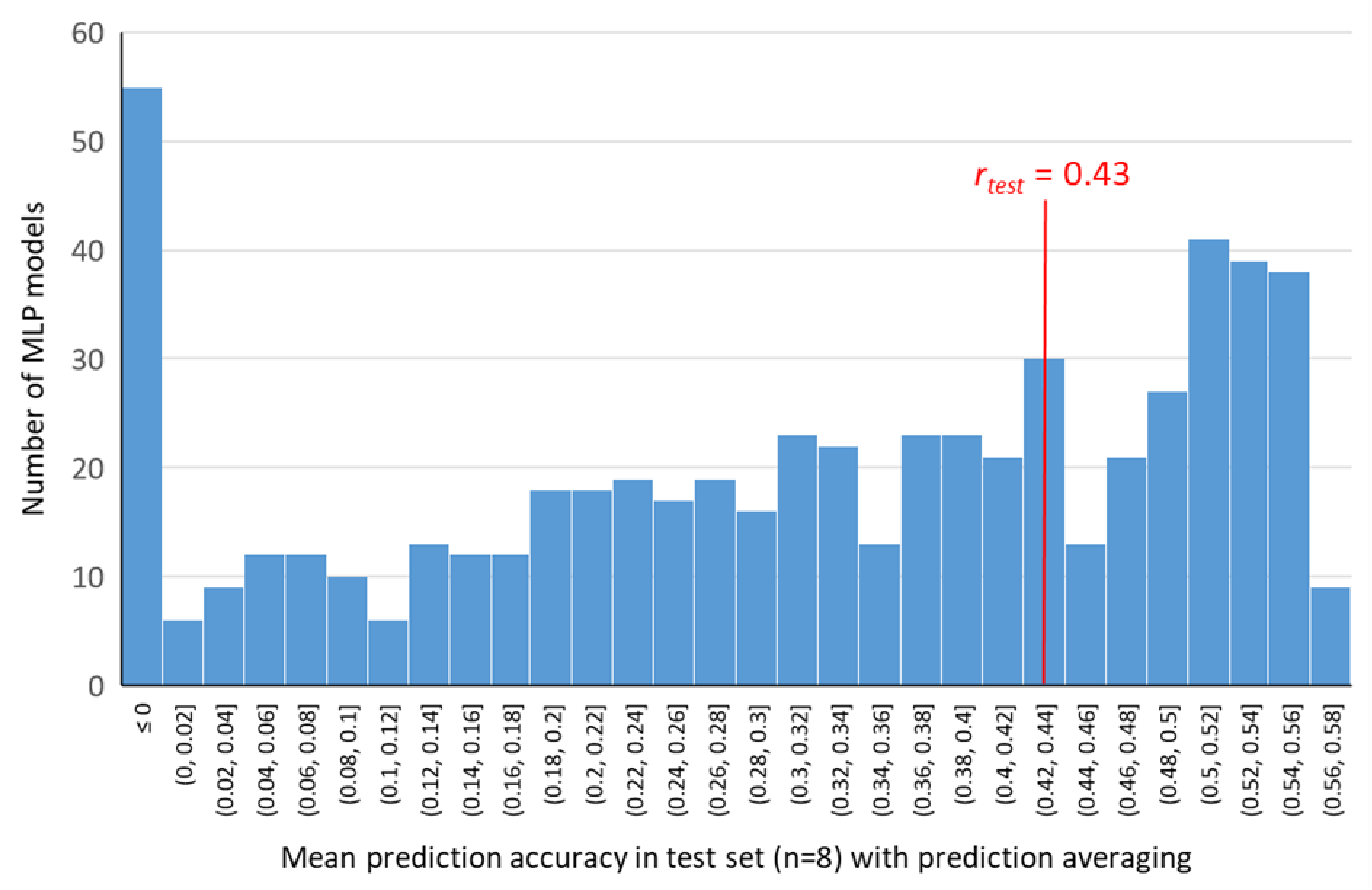
Practical application = **no test data available**

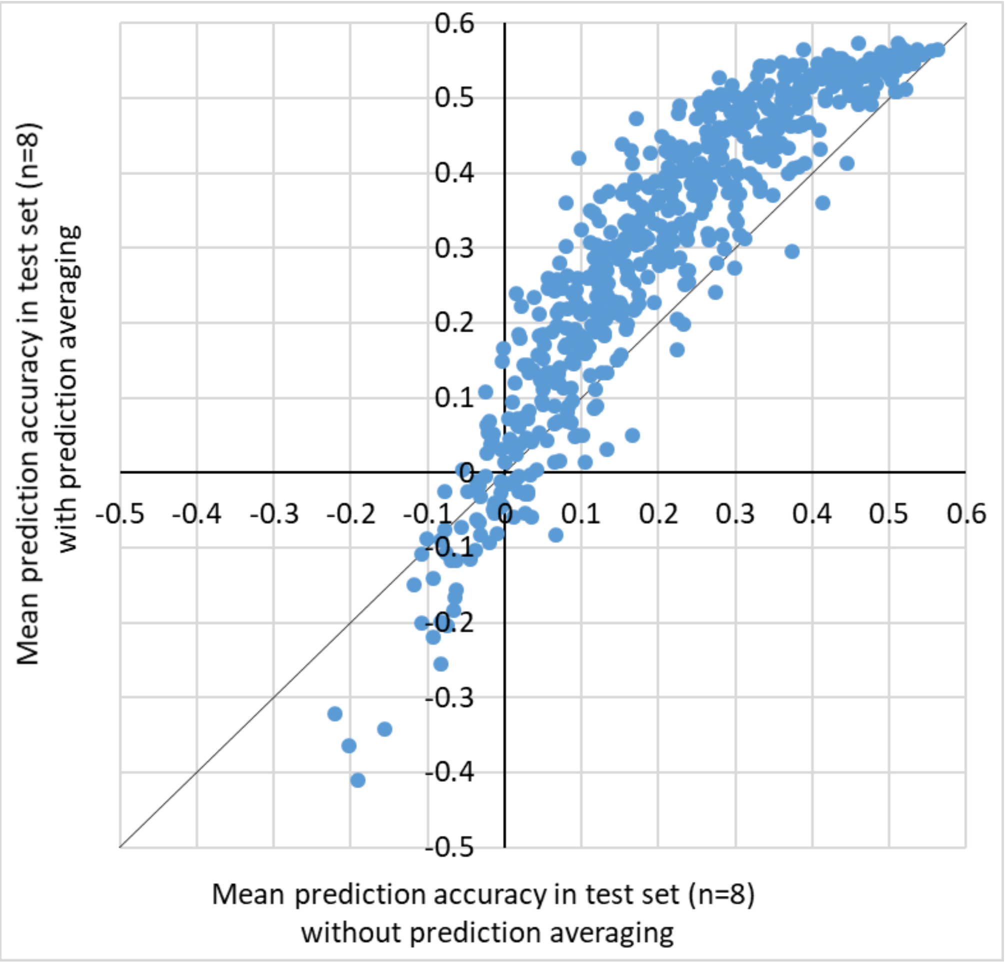
Question 4. How to optimize ANN using the training data ?

TensorFlow Keras

Question 1. What is the variability in r_{test} among MLP ?

Question 2. How to compute r_{test} : $\overline{cor(y,\hat{y})}$ ou $cor(y,\overline{\hat{y}})$?



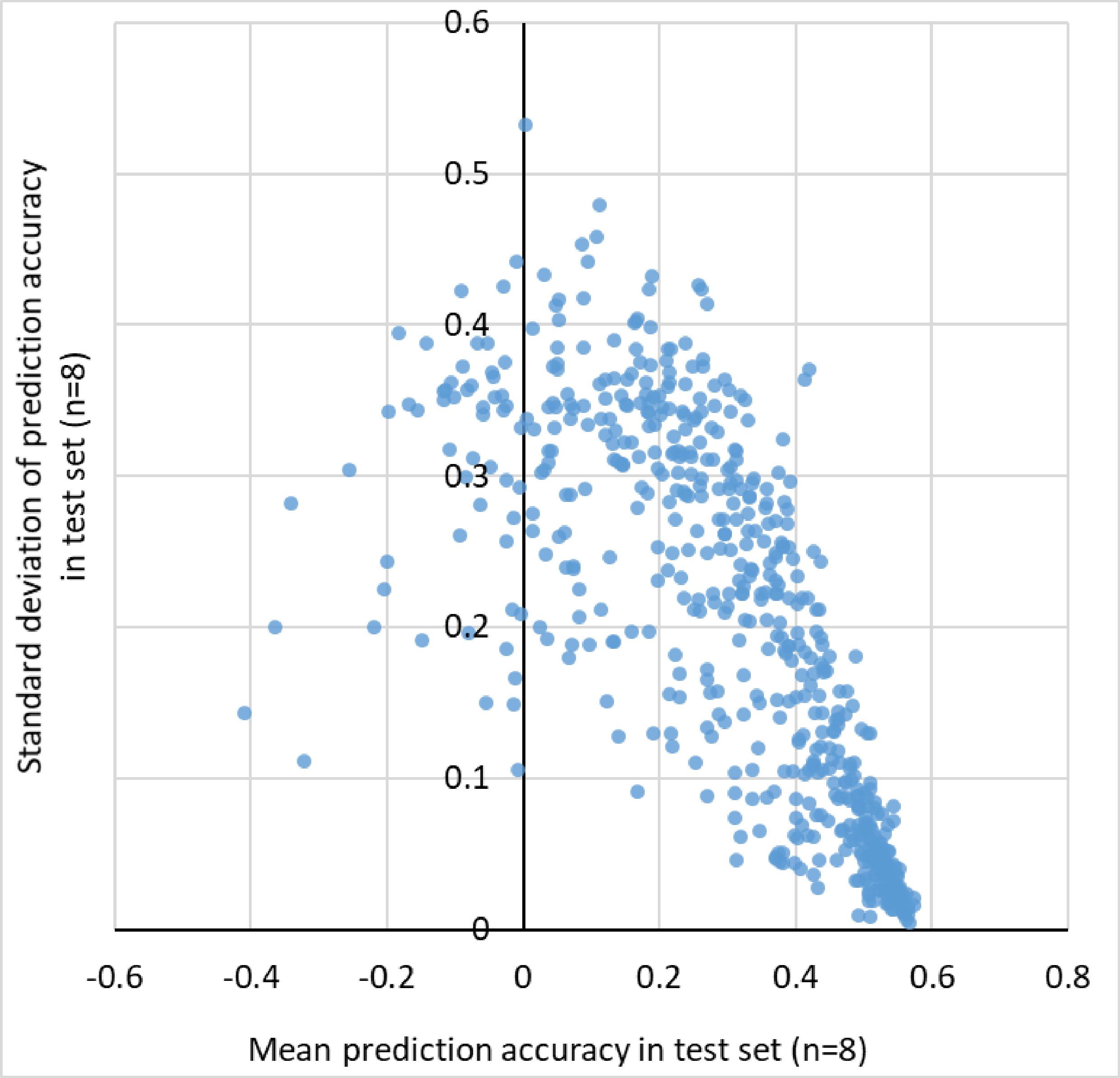


- **High variability in r_{test}** [-0.41,0.57]
- Many MLP models outperform conventional methods (35.6% of models with $r_{test} \geq 0.43$, **5% of MLP with $r_{test} \geq 0.55$**)

- **Prediction averaging increases r_{test}**

597 MLP models, validation subsets = optimized k-folds (k=5), r_{test} = means over 8 values

Question 3. What is the variability in r_{test} for a given MLP and dataset ?

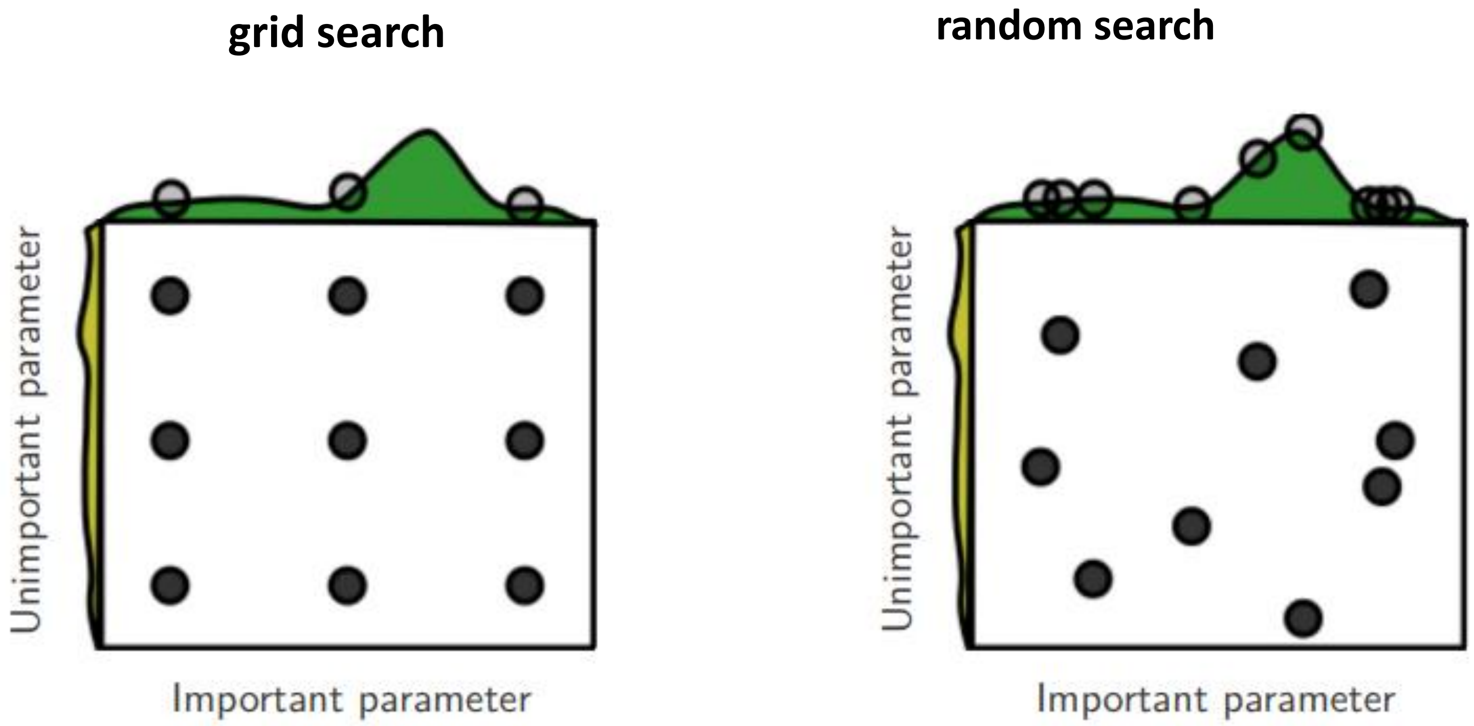


- **Model repeatability can be very low but is high for good models**
- **Good to make a few replicates to accurately identify best models**

597 MLP models, validation subsets = optimized k-folds (k=5), r_{test} = means over 8 values

Question 4. How to optimize ANN using the training data ?

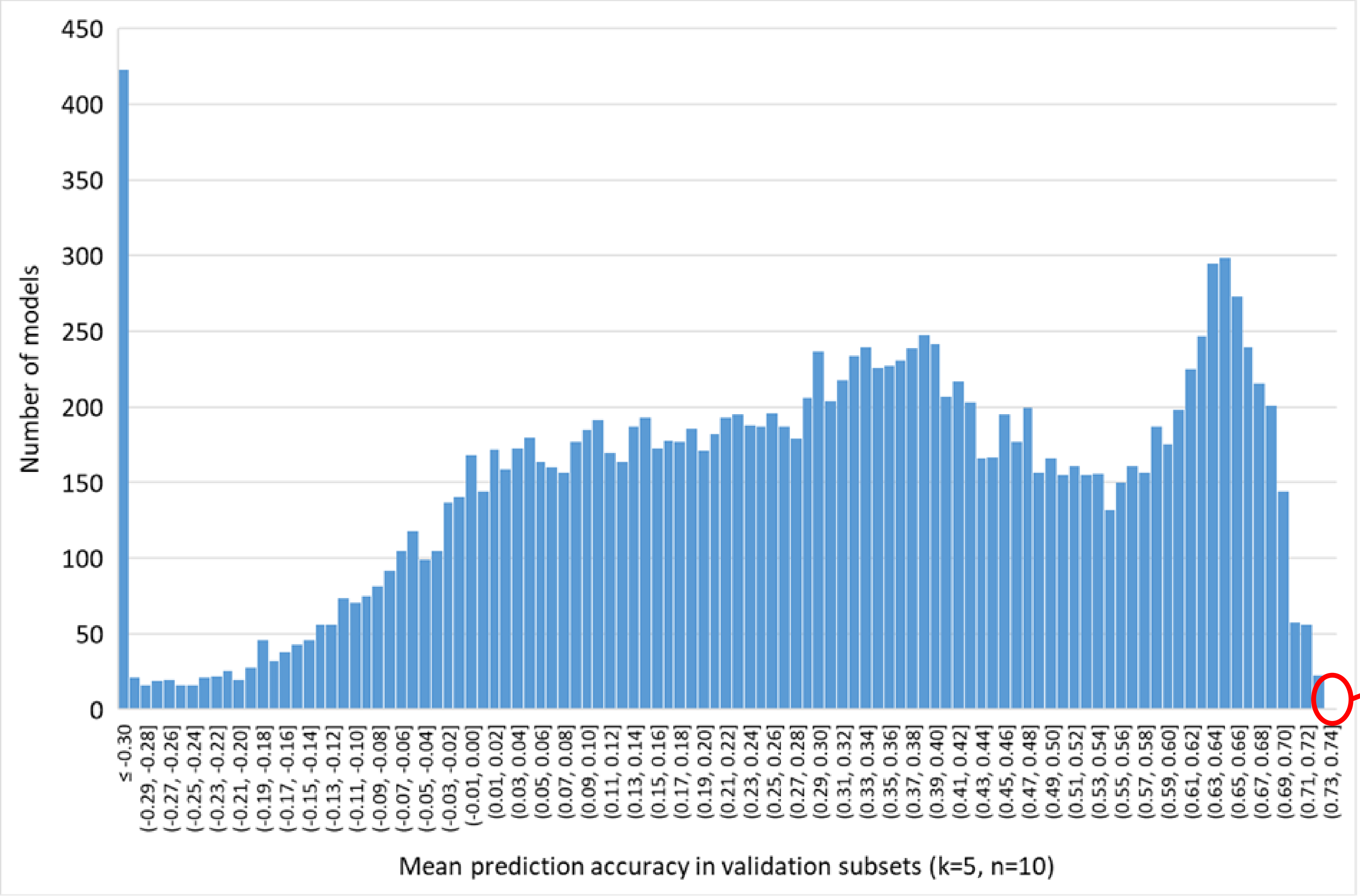
- **Different optimization methods developped:**



Bergstra and Bengio 2012

Question 4. How to optimize ANN using the training data ?

- **random search** for MLP, with 15874 random models:



Top 3 models on r_{val} :

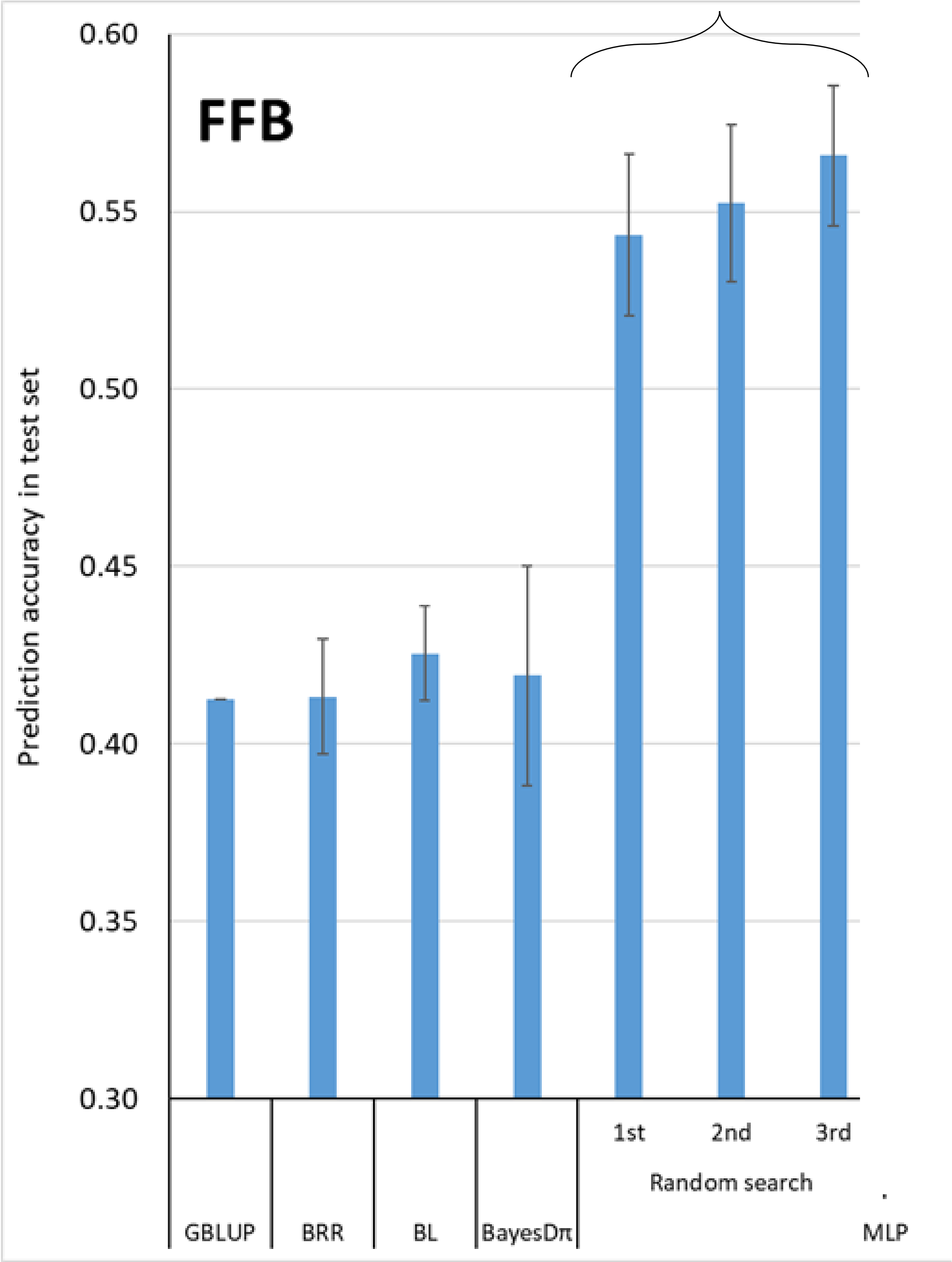
1st: 0.731

2^{nd}: 0.729

3rd: 0.728

Prediction accuracies in test set:

+ 27.6% to +32.8%



Question 4. How to optimize ANN using the training data ?

Random search = a lot of models to test

~16K here = **~22.8 days of GPU computing time**

64K in Sousa et al. (2022) [MLP for coffee], even more needed if space of architectures / hyper-parameters increases (in particular for more complex types of ANN) and/or if size of dataset increases

→ financial and GHG cost

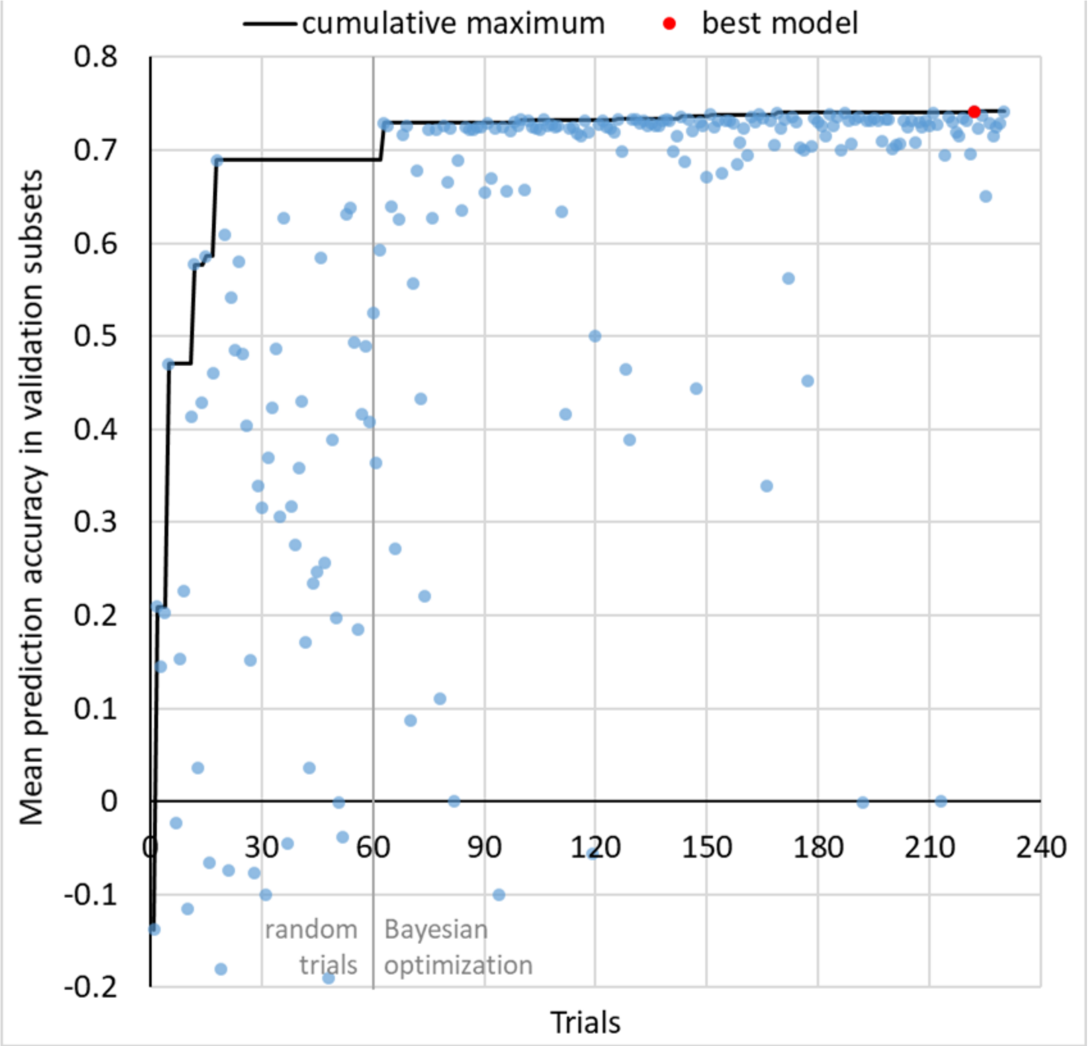
+ not guarantee to find the best model

→ **could we optimize models more efficiently (faster and/or to get higher GS accuracy) ?**

Bayesian optimization Iterative algorithm to uncover the global maxima of a black-box function in the defined parameter space

pyGPGO: Bayesian optimization for Python
(Jiménez and Ginebra 2017)

Example result (same range of architectures and hyper-parameters as random search):



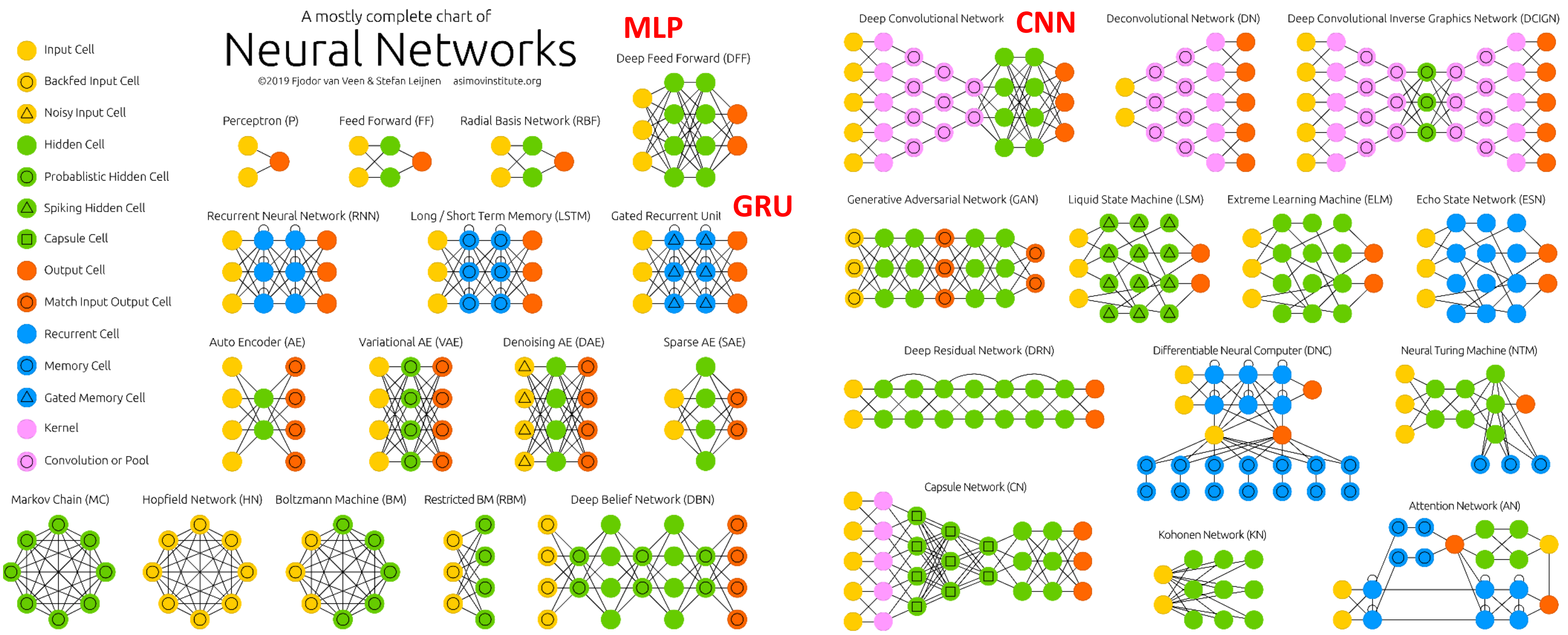
Number of models tested before r_{val} > random search = 73
(range 62-100, n=5)

→ Identify very fast models that outperform best model of random search

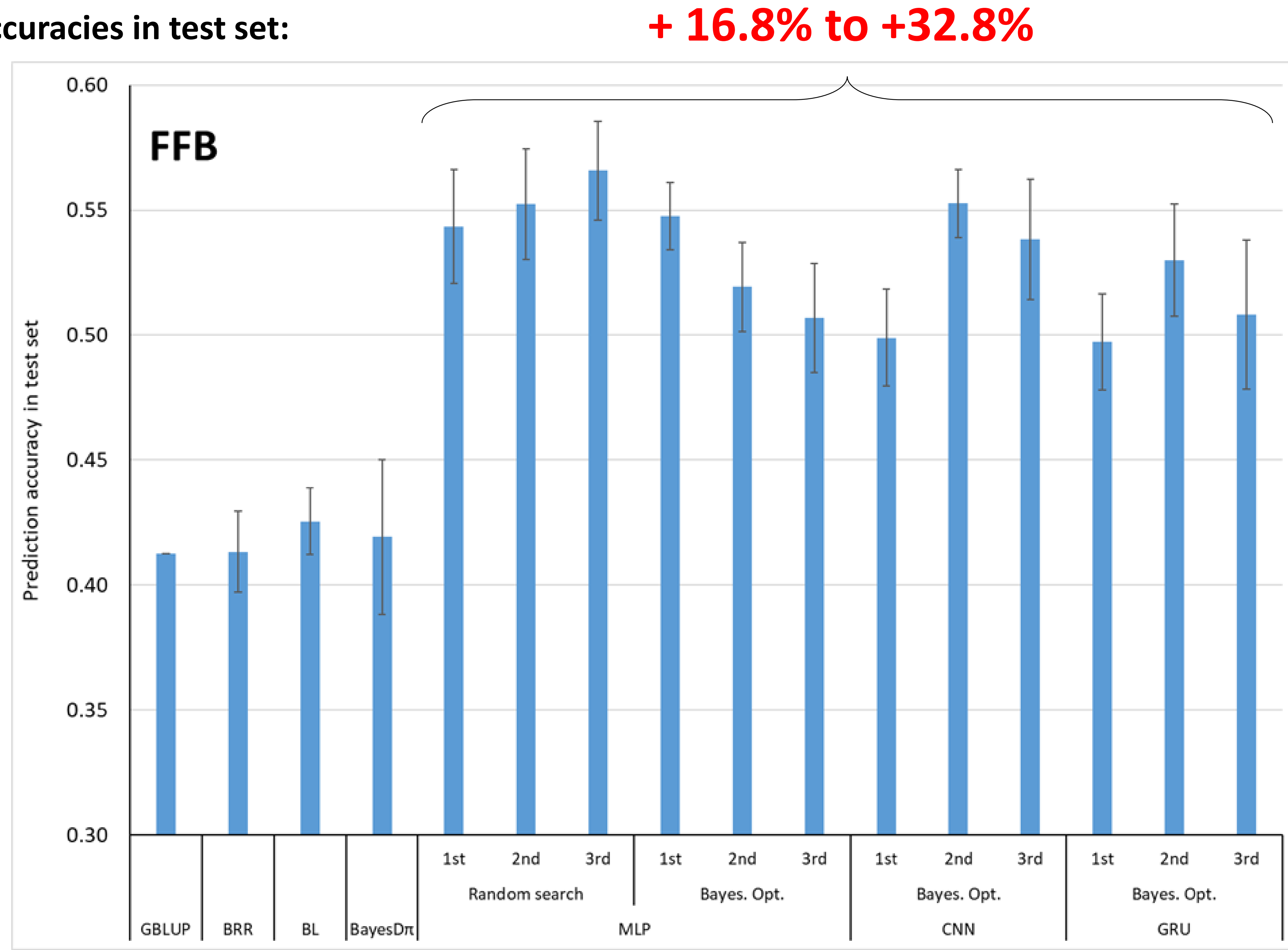
43 hours per run of Bayesian optimization (41.8-44.7)

Small gain in maximum r_{val} compared to random search: +1.16% (range 0.80-1.51), with on average 207 trials (165-248)

Beyond MLP - a lot of more complex models:



Prediction accuracies in test set:



Conclusions:

→ High variability in predictive ability depending on architecture/hyper-parameters of ANN models

→ High repeatability for good ANN models

→ Prediction averaging increases predictive ability in ANN

→ No effect of type of ANN (MLP, CNN, GRU)

→ Training data can be used to identify models giving large increases in GS accuracy

→ Bayesian optimization efficient to identify good ANN

Conclusions:

More details & results in:

Preprints, Working Papers, ... (Preprint) Year : 2025

HAL open science

Optimizing artificial neural network methodologies for enhanced genomic predictions: a case study with oil palm (Elaeis guineensis) data

David Cros (1) , Lauriane Rouan (1) , Daphné Navratil (1) , Billy Tchounke (2) , Nicolas Leroy (1) , Sandrine Le Squin (3) , Najelaa Ulfah (4) , Léifi Nodichao (5) , Grégory Beurier (1)

Show details

1 CIRAD, UMR AGAP Institut, F-34398 Montpellier, France
2 Department of Plant Biology, Faculty of Science, University of Yaoundé I, Yaoundé, Cameroon
3 PalmElit SAS, 34980 Montferrier sur Lez, France
4 P.T. SOCFINDO Medan, Medan, Indonesia
5 INRAB, CRA-PP, Pobè, Benin

→ Large trait effect: +5.1% in r_{test} for bunch number, same r_{test} for height increment

→ Computation time can be decreased further through complexity reduction methods

→ Contrasted ANN have similar prediction accuracy

→ Correlation between prediction accuracy in test subset and validation subsets is a key factor for the efficiency of model optimization

On-going / prospects:

- Multimodal approaches: SNP + weather data
- ANN model improvement
- Use of other machine learning approaches
- Multi-trait models

…

(we are hiring! - deadline Jan 19, 2025 ☺)

Researcher in deep learning to support plant improvement

Apply for vacancy

Thanks for your attention!