KI-Zucht: Künstliche Intelligenz als Basis für die auf “Big Data” basierte Pflanzenzüchtung der Zukunft

AI-based models and environmental data improves genomic selection in canola

Dr. Iulian Gabur

Department of Plant Breeding, Justus Liebig University Giessen, Germany

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Plant Breeding Giessen

Why A.I. in plant breeding?

Breeders equation

$$\Delta G = \frac{i * r * \delta_G}{L_g}$$

i = Selection intensity

r = Selection accuracy

δ_G = Genetic variance

L_g = Length of breeding cycle

Why A.I. in plant breeding?

Breeders equation

$$\Delta G = \frac{i * r * \delta_G}{L_g}$$

	How to improve ΔG?
i = Selection intensity	Select more intensely
r = Selection accuracy	Select more accurately
δ_G = Genetic variance	Introduce new variation
L_g = Length of breeding cycle	Do all faster!

Why A.I. in plant breeding?

Breeders equation

$$\Delta G = \frac{i * r * \delta_G}{L_g}$$

	How to improve ΔG?	ML models?
i = Selection intensity	Select more intensely	Digital phenotyping, drones (MLP, NN)
r = Selection accuracy	Select more accurately	3D plant phenomics (FS, RF, CNN)
δ_G = Genetic variance	Introduce new variation	Conserve/increase diversity (GA, NLP)
L_g = Length of breeding cycle	Do all faster!	Genomic selection (SVM, GBM, NN)

Why A.I. in plant breeding?

Breeders equation

$$\Delta G = \frac{i * r * \delta_G}{L_g}$$

KEYNOTE TALK

#034 Omics-based optimisation of hybrid performance and heterosis in winter oilseed rape

Rod Snowdon, Sven Weber, HueyTyng Lee, Agnieszka Golicz, José Montero, Mauricio Orantes-Bonilla, Lennard Ehrig, Iulian Gabur, Eva Herzog, Amine Abbadi, Matthias Frisch

#093 Connecting rapeseed plant architecture to yield via 3D-imaging

Sven Weber, Andreas Eckert, Lennard Ehrig, Rod Snowdon, Andreas Stahl

#049 HaploMAGIC: phasing and accurately detecting recombination in multi-parental populations with genotyping errors

Jose Antonio Montero-Tena, Nayyer Abdollahi-Sisi, Amine Abbadi, Matthias Frisch, Rod J. Snowdon, Agnieszka Golicz

#074 Unravelling the *Brassica napus* epigenetic network with an integrated multi-omics approach

Silvia Zanini, Rod Snowdon, Agnieszka Golicz

#073 Phenomic selection: predicting quantitative traits in canola with NIRS-profiles

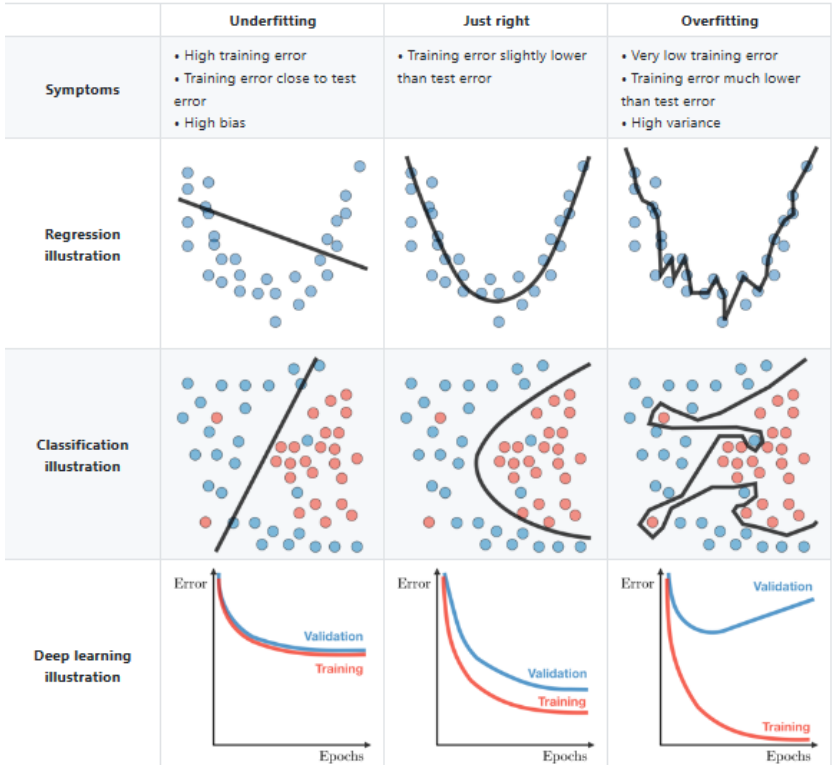
Lennard Ehrig, Sven Weber, Stefan Abel, Reinhard Hemker, Dirk Stulgies, Milka Malenica, Amine Abbadi, Benjamin Wittkop, Rod Snowdon, Andreas Stahl

Reducing dimensionality in big-data

Idea:
Identify the most important SNP for each Trait-SNP association (**feature selection**) and use them for training and predictions.

Avoid the **curse of dimensionality** and **overfitting** !

two wind turbines in two dimensional and three dimensional space



Using linear and non-linear models in canola genomic selection

Idea:
Identify the most important SNP for each Trait-SNP association (**feature selection**) and use them for training and predictions.

Testing:

Used **three** linear and non-linear approaches for **feature selection**:

- **PCA -** data reduction (100 features)
- **RF-based** reduction (1000 features)
- Random selection of features (1000 features)

Statistical models used in a cross validation pipeline:

- Ridge-Regression Best Linear Unbiased Prediction (RR-BLUP)
- Least Absolute Shrinkage and Selection Operator (LASSO)
- "general" Linear Model (GLM)
- **Gradient Boosting Machine (GBM)**
- **Artificial Neural Network (ANN)**
- **Random forest (RF)**

Hybrid canola dataset

PREDICT data set – spring-type *B. napus* were evaluated at four different locations across Denmark, Germany, Poland and Estonia. Jan et al., 2016 PLoS ONE 11(1): e0147769.

Design:
two male sterile testers (MSL-T1 and MSL-T2, NPZ Lembke, Hohenlieth, Germany) and a diverse population of 475 spring-type “00” *B. napus* cultivars = 950 hybrids

SNP genotyping
Brassica 60K Illumina Infinium™ SNP genotyping array of parental lines

Population
950 F1 hybrids – *in silico* crossing scheme (0, 1, 2)
Werner et al. 2018 Theor Appl Genet 131, 299–317

Phenotype data of hybrids (14) :
Yield [dt/ha],Oil-Yield [dt/ha],Moisture [%],Oil [%],Protein Seed [%],Protein Meal [%]
GSL Meal ymol/g seed,GSL ymol/g seed,TSW [g], Days to flower, Straw length [cm]

Workflow

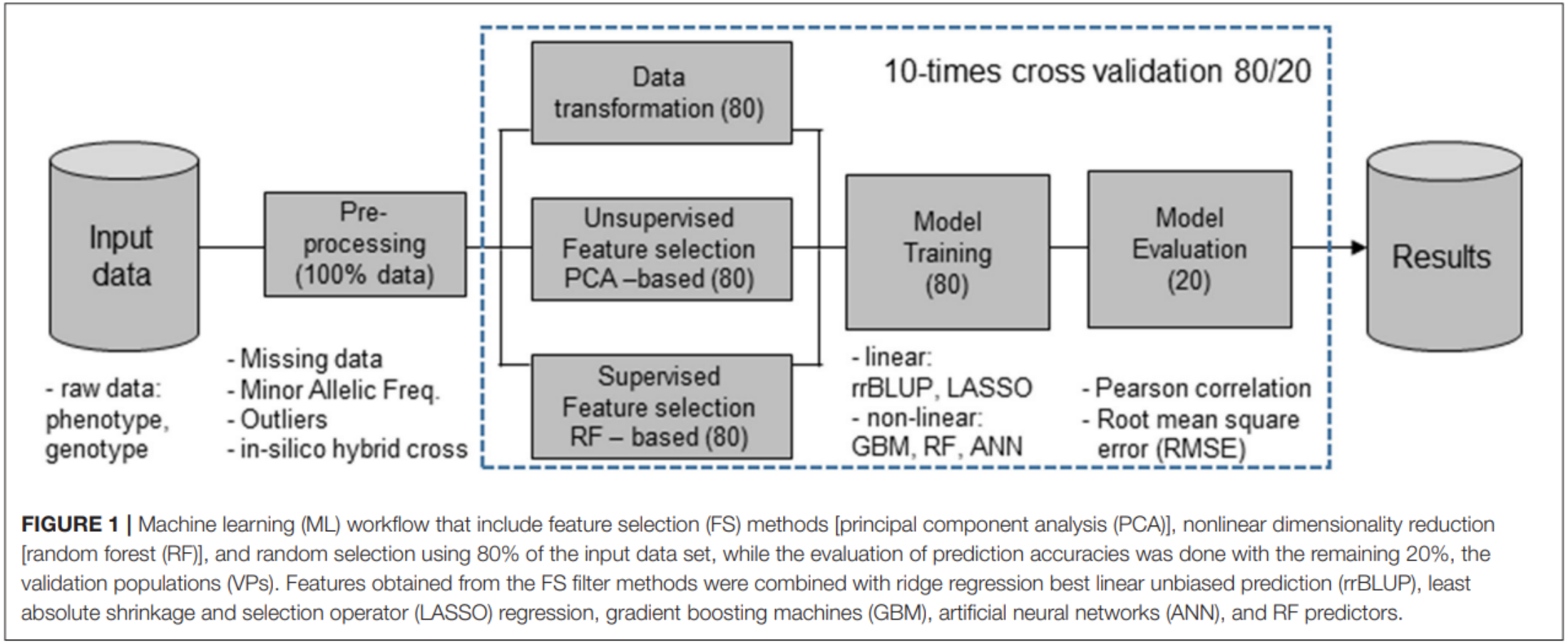


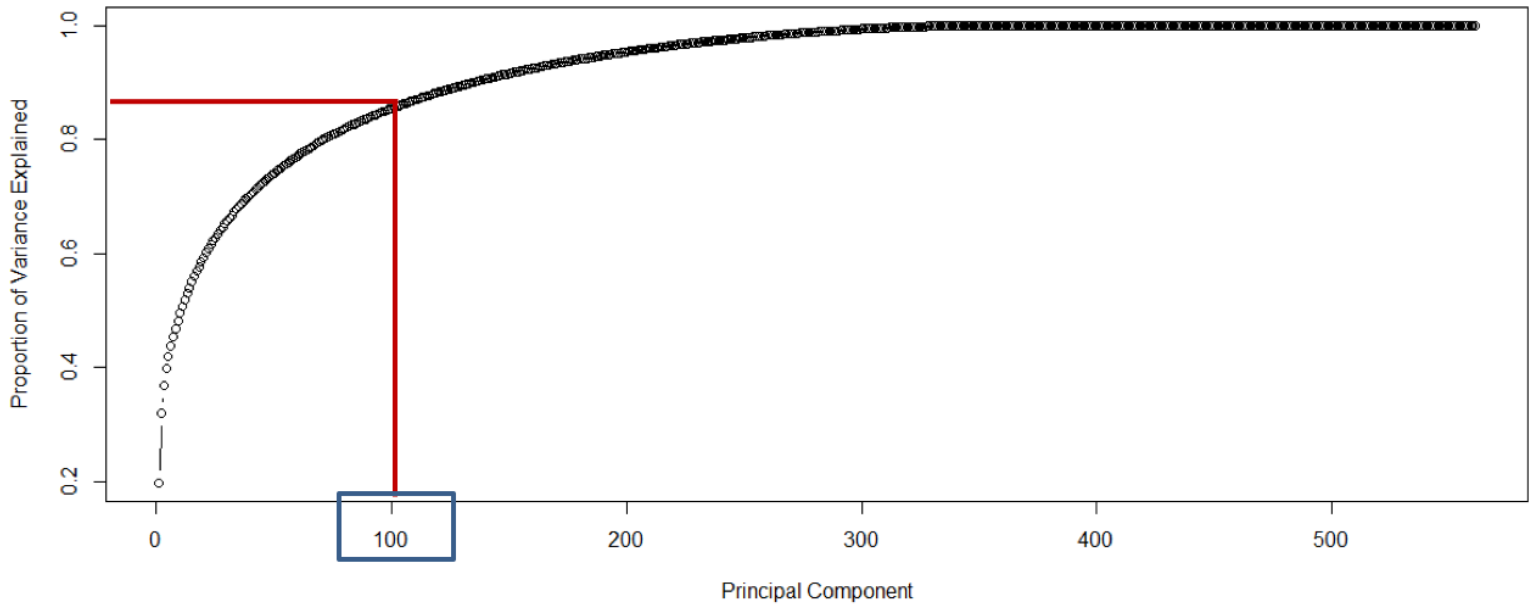
FIGURE 1 | Machine learning (ML) workflow that include feature selection (FS) methods [principal component analysis (PCA)], nonlinear dimensionality reduction [random forest (RF)], and random selection using 80% of the input data set, while the evaluation of prediction accuracies was done with the remaining 20%, the validation populations (VPs). Features obtained from the FS filter methods were combined with ridge regression best linear unbiased prediction (rrBLUP), least absolute shrinkage and selection operator (LASSO) regression, gradient boosting machines (GBM), artificial neural networks (ANN), and RF predictors.

Gabur et al (2022) Front. Artif. Intell. 5:876578

Top 100 PCs explain 87% of genetic variation in the canola dataset

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_d], \quad \boldsymbol{x} \in \mathbb{R}^d$$
$$\downarrow \boldsymbol{x}\boldsymbol{W}, \quad \boldsymbol{W} \in \mathbb{R}^{d \times k}$$
$$\boldsymbol{z} = [z_1, z_2, \ldots, z_k], \quad \boldsymbol{z} \in \mathbb{R}^k$$

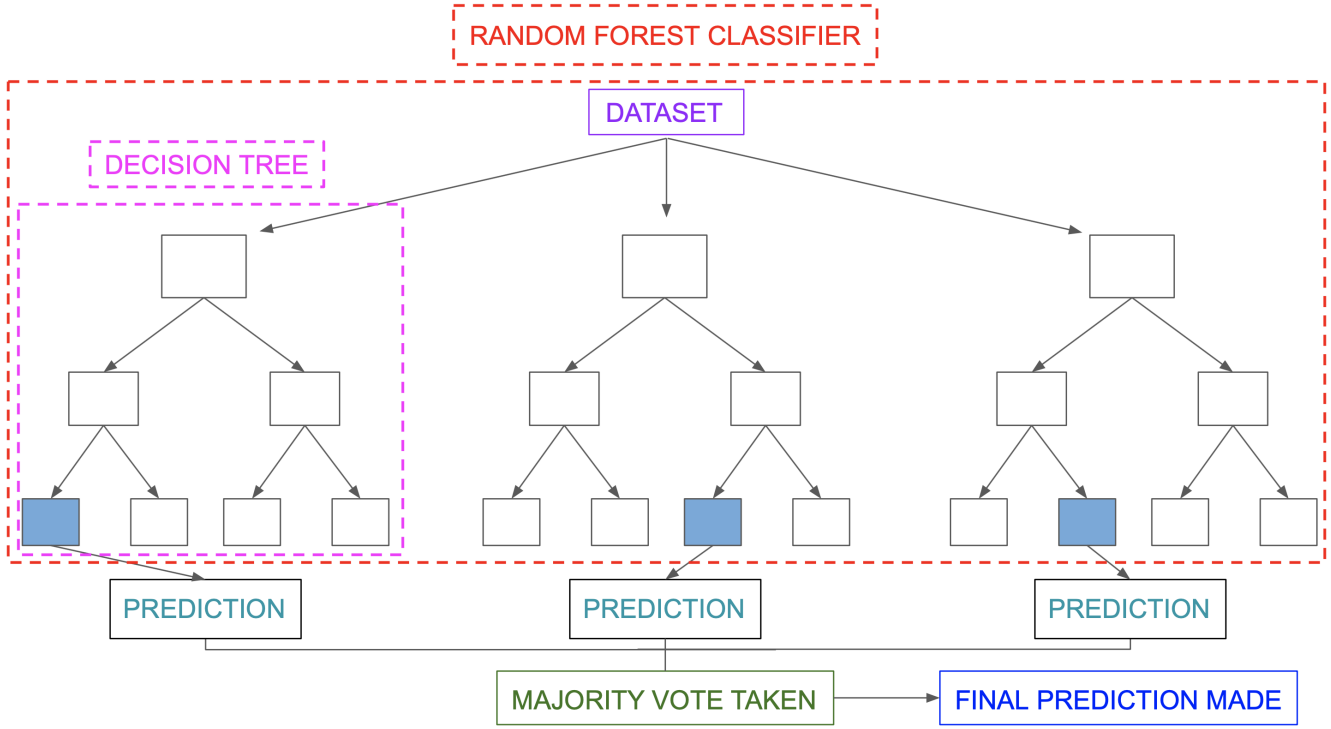
1. Select **k** eigenvectors for k largest eigenvalues (k ≤ d).
2. Construct **matrix W** from the "top" k eigenvectors.
3. Transform the d-dimensional input **dataset X** using the projection **matrix W** to obtain the new k-dimensional feature subspace.



Gabur et al (2022) Front. Artif. Intell. 5:876578

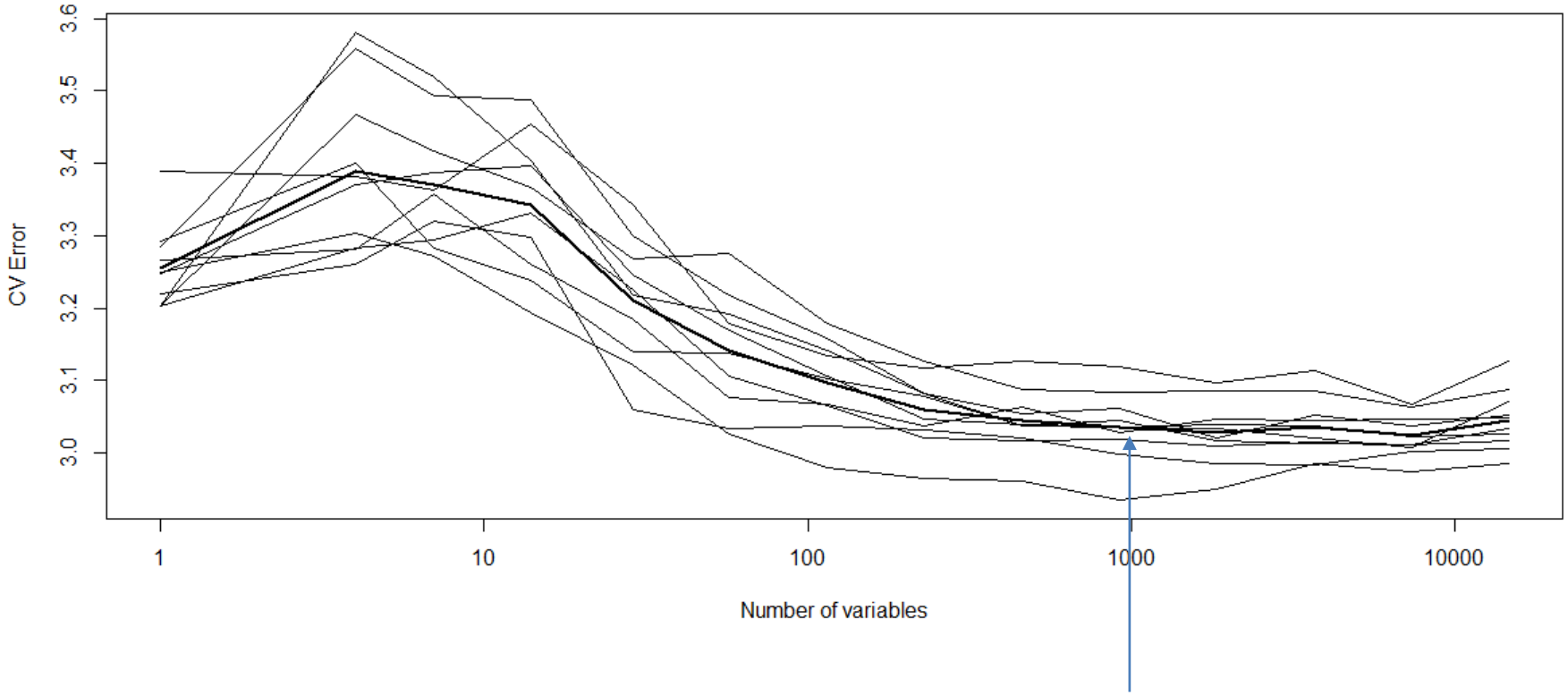
Random forest (RF) based data reduction for identification of top 1000 features

Identify the maximum variance in SNP matrix and calculate **feature importance** scores for each feature based on the ‘Gini’ criterion (RF trees).



Siddharth Misra, Yaokun Wu, in *Machine Learning for Subsurface Characterization*, 2020

1000 trees reduce the error rate of Random Forest model

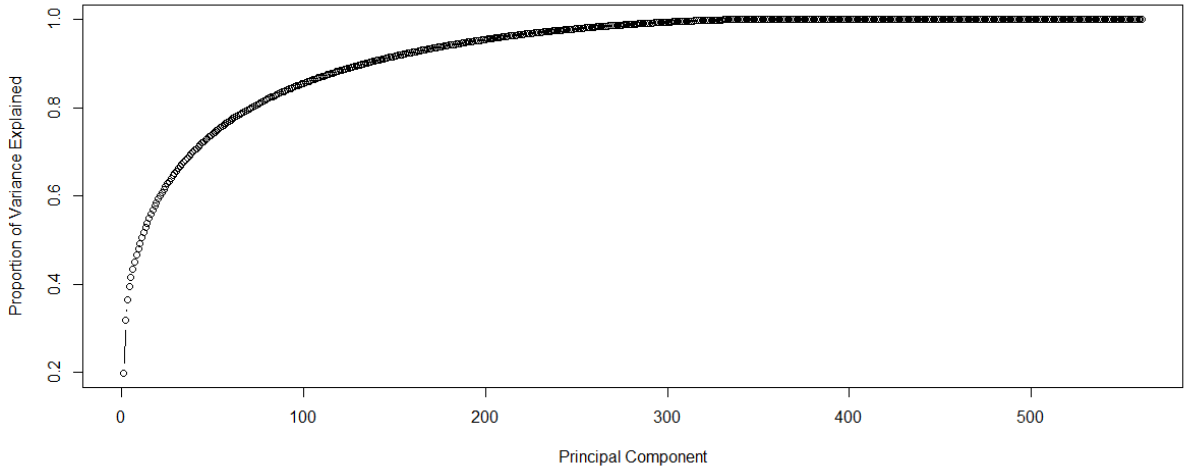


Select 1000 trees for predictions

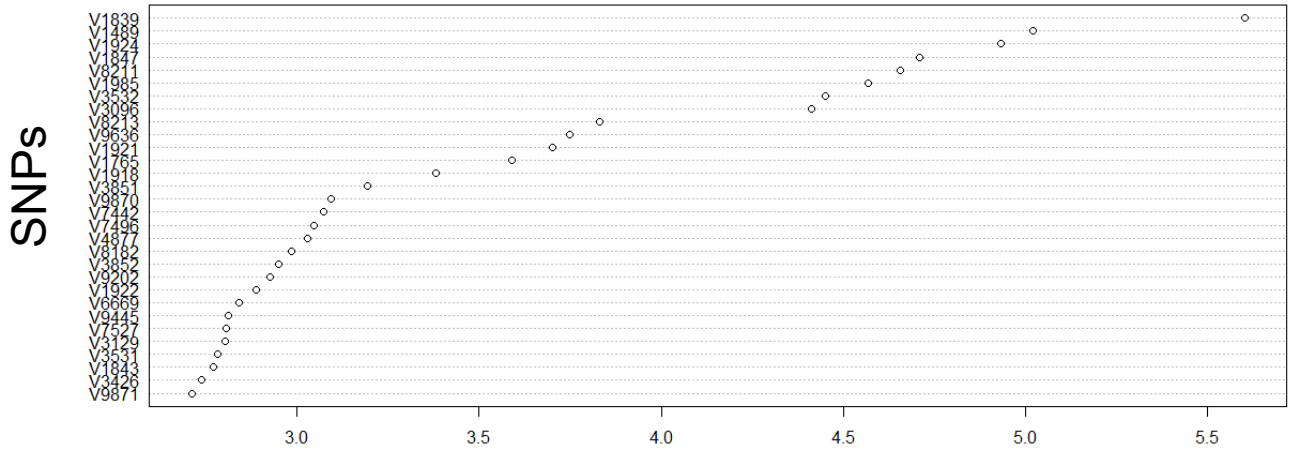
Gabur et al (2022) Front. Artif. Intell. 5:876578

Random Forest – based selection using features importance (top 1000)

PCA-based selection



RF selection

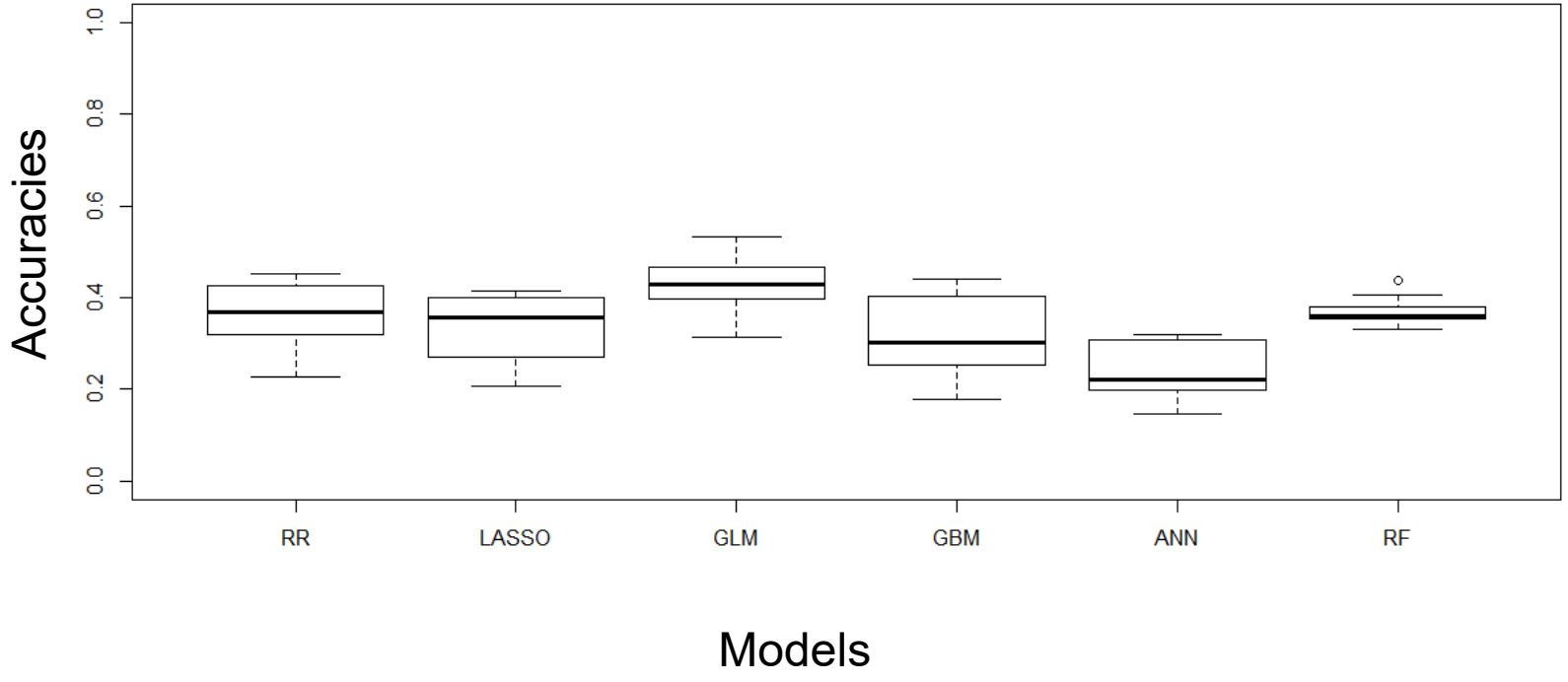


Feature importance /‘Gini’ criterion

Predictions using the entire canola dataset are time consuming

Phenotypic trait : **hybrid yield**
Models: RR-BLUP, LASSO, GLM, GBM, ANN, RF
Cross-validation: 10x – 80%/20%

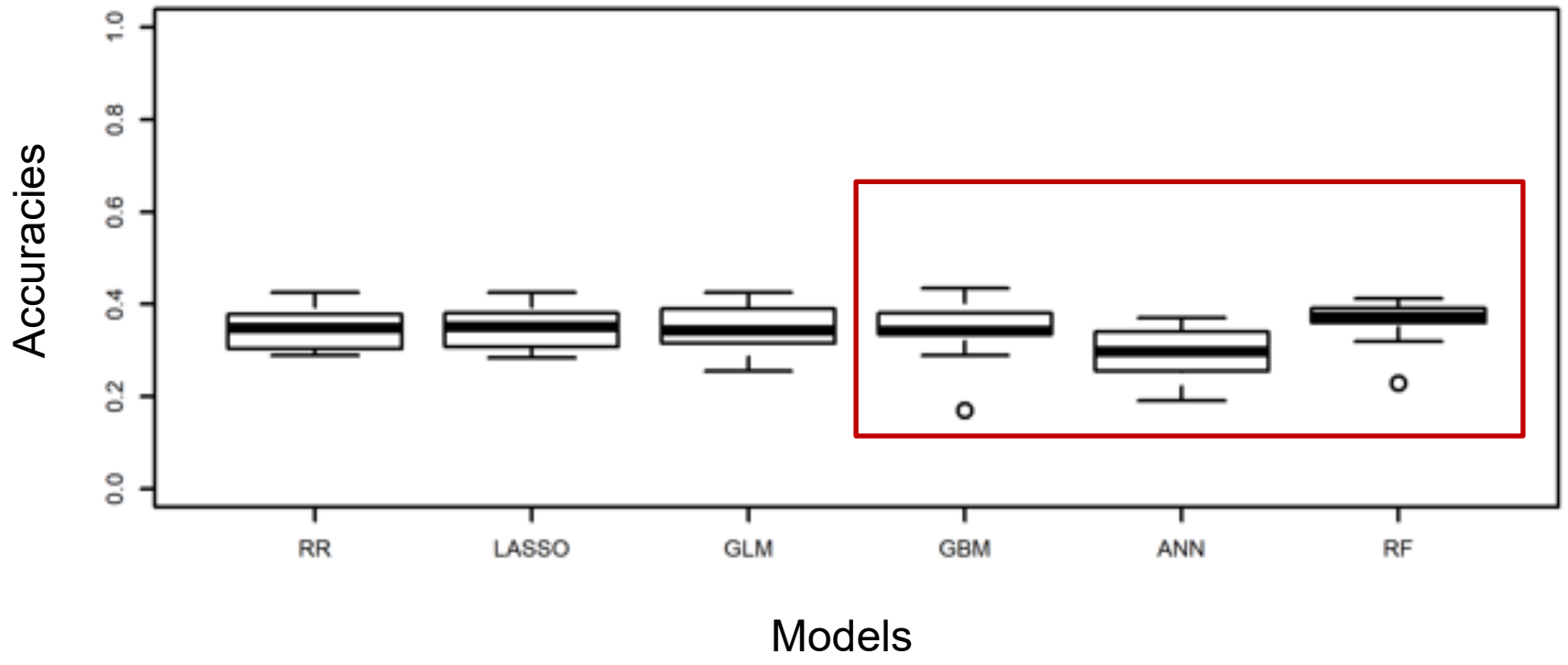
Entire set prediction – **approx. 1200 min/20h**



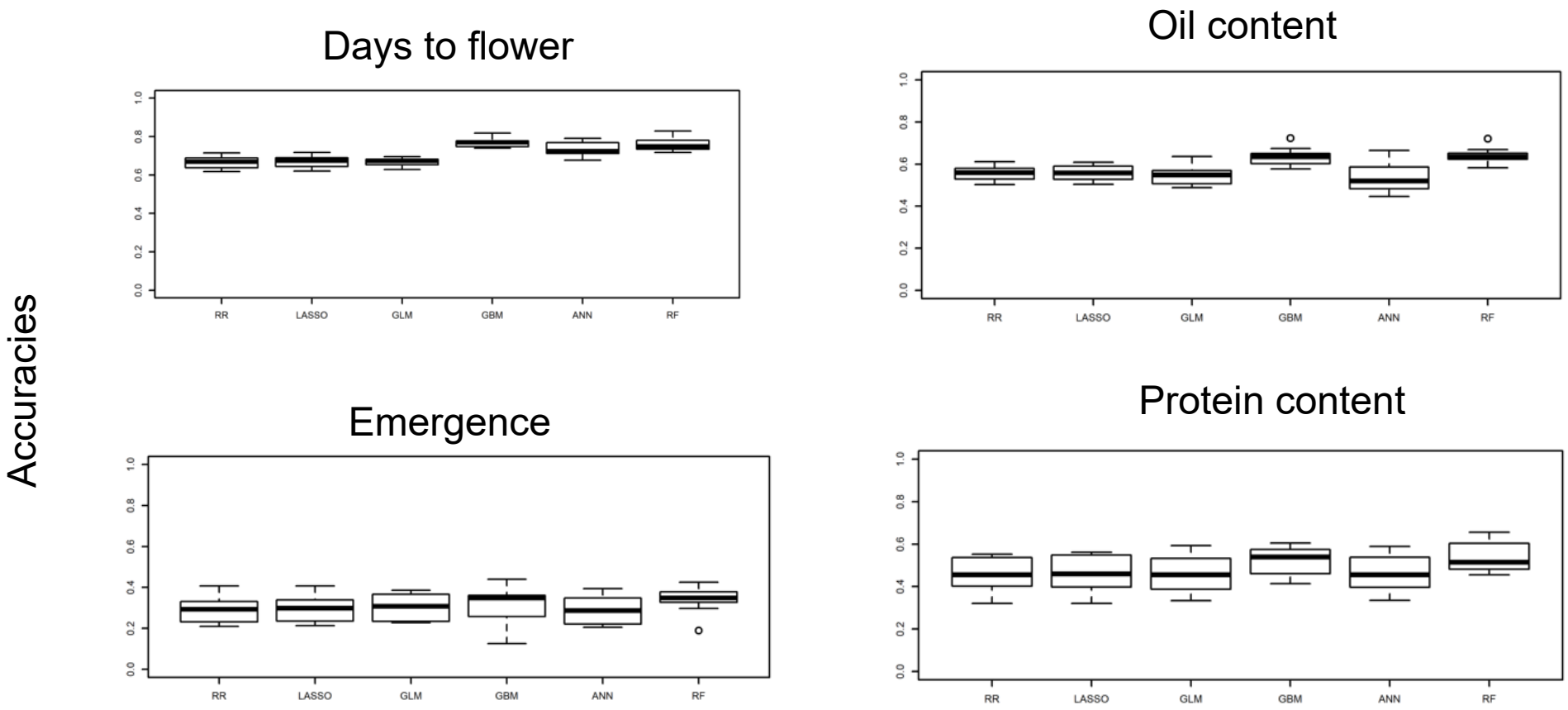
Predictions using the PCA-based FS are 200x faster !

Phenotypic trait : **hybrid yield**
Models: RR-BLUP, LASSO, GLM, GBM, ANN, RF
Cross-validation: 10x – 80%/20%
Matrix : **first 100PCs**

Entire set prediction – **approx. – 6.7 min** (200 x faster then the SNP matrix X)



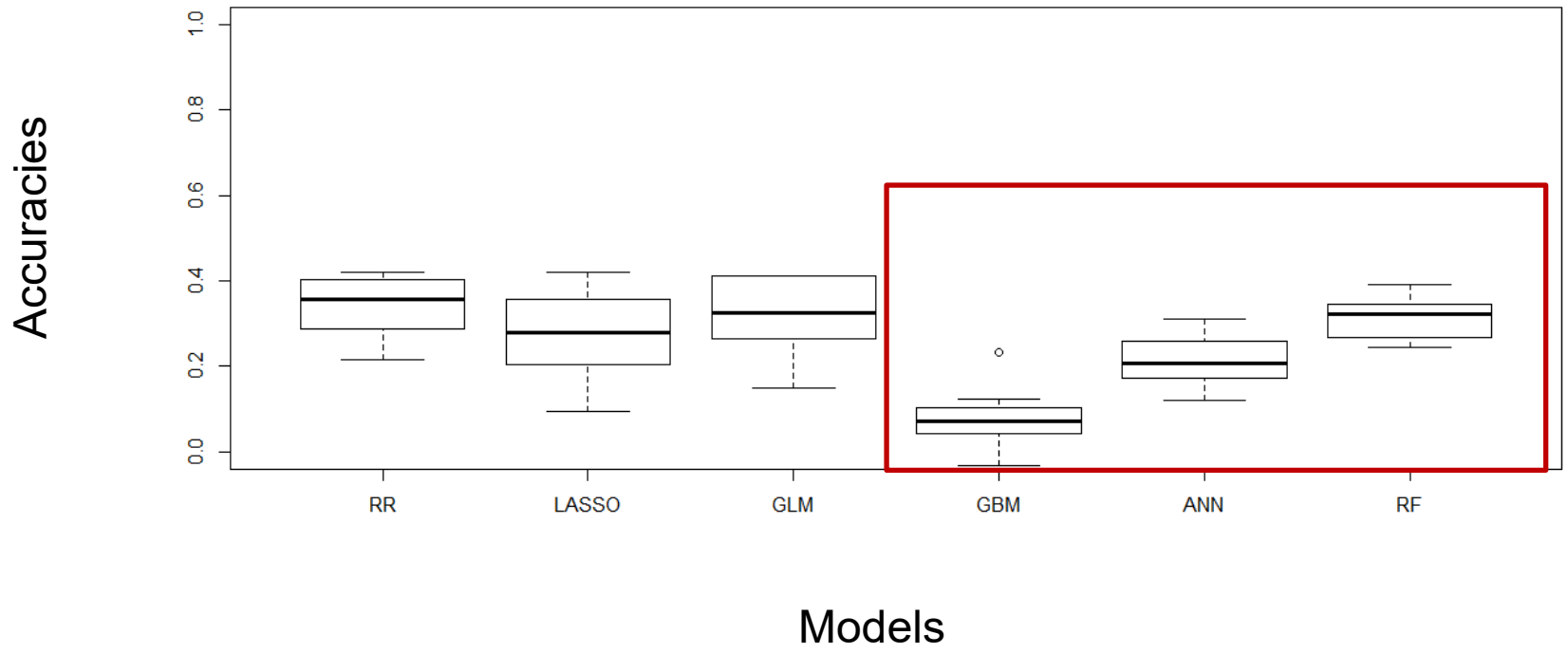
Similar results for all investigated traits using PCA-based FS



Predictions using the RF-based FS are 10x faster !

Phenotypic trait : **hybrid yield**
Models: RR-BLUP, LASSO, GLM, GBM, ANN, RF
Cross-validation: 10x – 80%/20%
Matrix : **first 1000 RF features**

Entire set prediction – **approx. – 119.6 min** (10 x faster then the SNP matrix X)



Feature selection improves prediction accuracies of non linear models

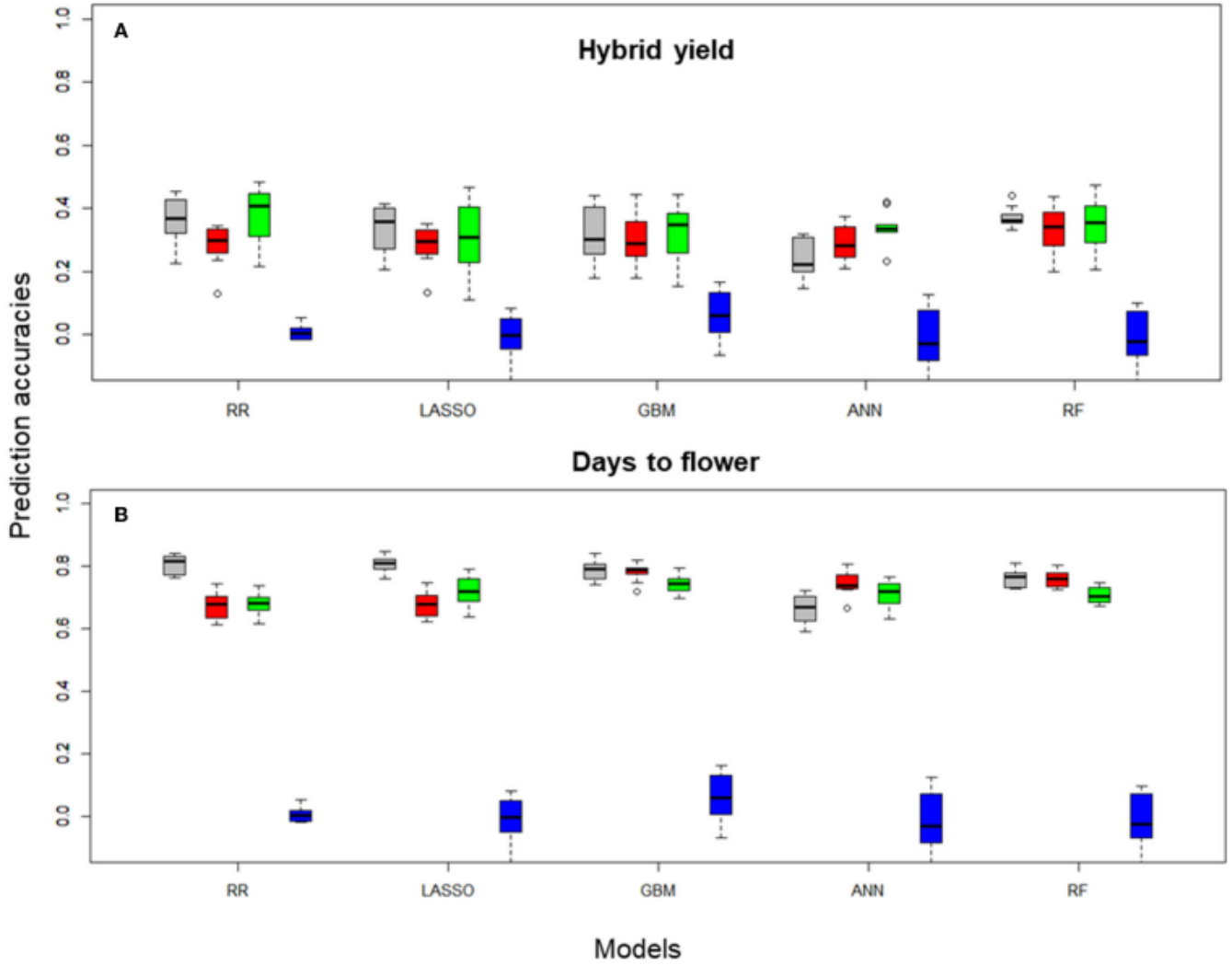


FIGURE 2 | Boxplots for **(A)** hybrid yield and **(B)** days to flower predictions of the *B. napus* data set obtained on 10-fold cross-validation (CV) sets, using Pearson correlations, with rrBLUP, LASSO, GBM, ANN, and RF, with 100, and 1,000 SNP subsets selected with different filter methods and the entire SNP data set (colored in gray). Filter methods: principal component analysis 100 PCA (red), 1,000 RF (green), 1,000 RS (blue), and 14,718 (gray, no FS using the total number of markers after QC).

Gabur et al (2022) Front. Artif. Intell. 5:876578

Feature selection improves prediction accuracies of non linear models

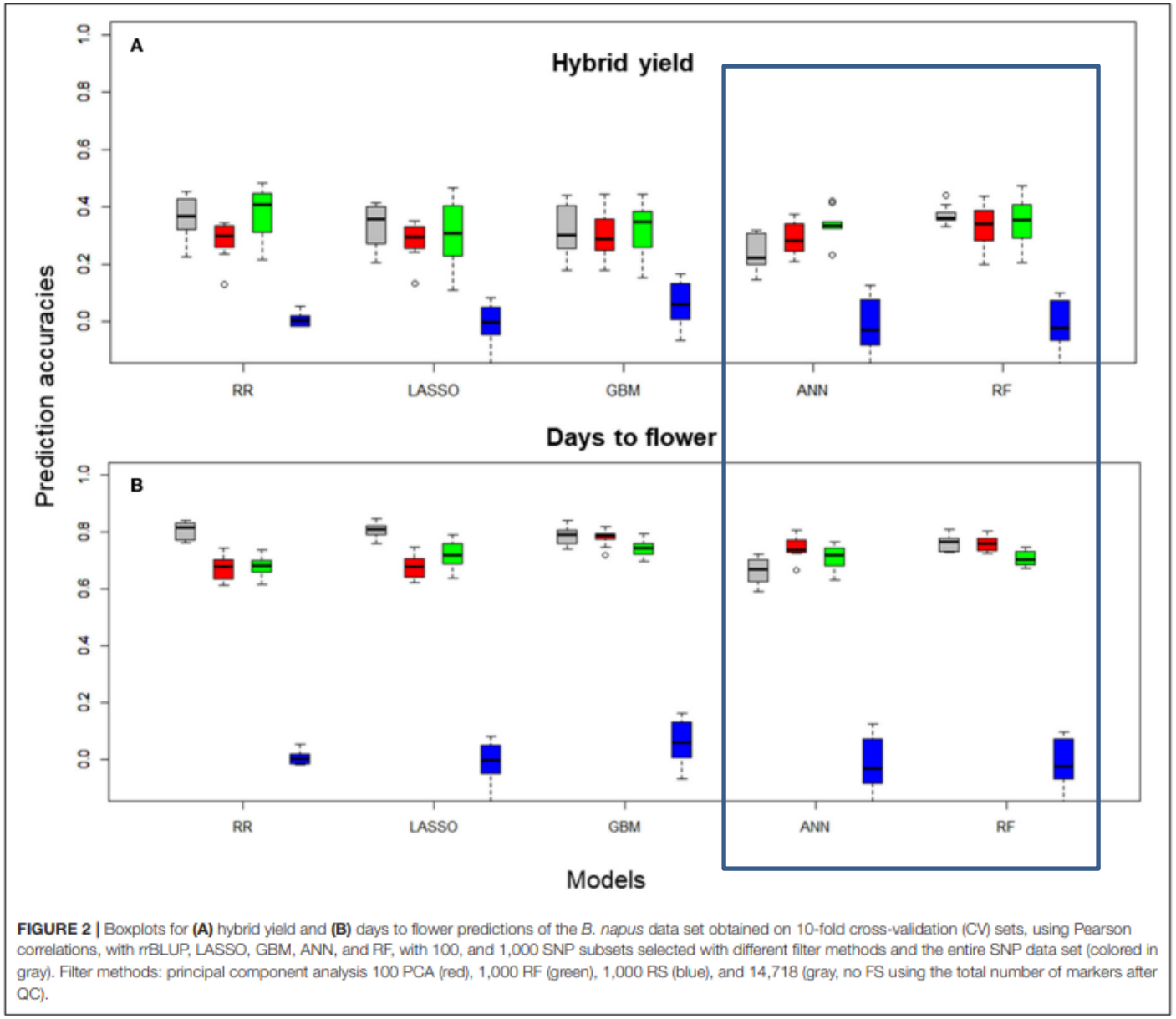


FIGURE 2 | Boxplots for **(A)** hybrid yield and **(B)** days to flower predictions of the *B. napus* data set obtained on 10-fold cross-validation (CV) sets, using Pearson correlations, with rrBLUP, LASSO, GBM, ANN, and RF, with 100, and 1,000 SNP subsets selected with different filter methods and the entire SNP data set (colored in gray). Filter methods: principal component analysis 100 PCA (red), 1,000 RF (green), 1,000 RS (blue), and 14,718 (gray, no FS using the total number of markers after QC).

Gabur et al (2022) Front. Artif. Intell. 5:876578

Acknowledgements

Justus Liebig University, Germany:
Prof. Rod Snowdon
Prof. Mathias Frisch
Dr. Christian Obermeier
Lennard Ehrig
Sven Weber

Faculty of Computer Science, UAIC-Iasi, Romania:
Prof. dr. Dan Cristea
Asso. Prof. dr. Mihaela Elena Breaban
Lector dr. Ionut Pistol
Asso. Prof. dr. Madalina Raship

Department of Computer Science, Alto University, Finland:

Institut of Bioinformatics, Kyoto University, Japan:
FiDiPro Prof. Hiroshi Mamitsuka

NPZ Innovation GmbH (NPZi), Germany:
Dr. Amine Abadi

All colleagues and technical assistants