#013

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

***Iulian Gabur*[1]**

Lennard Ehrig[1]
Rod Snowdon[1]

[1] Justus Liebig University, Giessen, Germany

Background:

Canola breeding is one of the economic sectors with the greatest productivity gains in the last years, with faster crop breeding progress of improved new varieties. Shaping the future of canola breeding requires that science and industry must concern themselves more closely with AI and agree on an agile, strategic approach.

Objective:

By better using available "big data" from modern breeding processes, AI models hold enormous potential to accelerate crop breeding progress for yield performance, more precisely manage field production and obtain information on the crop stability in different environments.

Methods:

Data from canola breeding programmes is generally multi-dimensional and lacks orthogonality, making it difficult to link different datasets. Furthermore, in most cases insufficient data volumes are available for individual breeders to effectively develop and train AI models. These challenges require completely new approaches for designing accelerated and cost-effective breeding processes. In order to exploit the potential of AI for practical breeding, it is essential to develop dedicated AI techniques that adapt the mathematical principles of AI to suit breeding purposes. Modern mathematical methods are needed to reduce the excess dimensionality of the data based on core attributes relevant for the desired predictions (so-called "feature selection", FS) or to account for effects of overfitting.

Results:

Recent advances in AI algorithms have enabled the construction of complex prediction models that can better discriminate between positive alleles and genetic background. Using data from canola breeding programs, we found that AI methods outperformed current state-of-the-art approaches, increasing prediction accuracy, significantly reducing computational time, and identifying key genetic regions involved in qualitative or quantitative traits.

Here we investigated linear and non-linear approaches for efficient FS methods in combination with parametric learners to predict important agronomical traits in a hybrid canola breeding population.

Predictive accuracy was generated by cross-validation for each FS method using appropriate input data and based on individual characteristics. Results for multiple models including rrBLUP, LASSO, GBM, and ANN suggest that nonlinear learners (GBM, ANN, and RF) tend to outperform linear learners (rrBLUP and LASSO) when the selection of relevant features is introduced in predictions across training sets. Tree-based methods (GBM and FR) exhibited a high prediction accuracy when combined with PCA-based dimensionality reduction and Random Forest-FS of the input data.

Conclusions:

AI-based prediction methods are likely to become key tools for applied plant breeders. Feature Selection complemented by modern nonlinear learners demonstrate the power to integrate large disparate datasets, enhance tools for greater predictive accuracy, and extract relevant functional meanings from genotype-phenotype relationships.

Our results demonstrate that significantly reducing the dimensionality of datasets helps toi significantly reduce computation time while improving prediction accuracy, especially for nonlinear learners. Efficient combinations of trait identification methods and modern predictive models can provide breeding programs with the tools they need to effectively use cost-effective genotyping data and improve predictive accuracy.