Using Machine Learning methods in our research at CiBreed

Researcher: Selina Klees
Paper: In Silico Identification of the Complex Interplay between Regulatory SNPs, Transcription Factors, and Their Related Genes in Brassica napus L. Using Multi-Omics Data
Link: https://doi.org/10.3390/ijms22020789

Another researcher we talked with is Selina Klees, who used Machine Learning Algorithms in order to combining the genomic, transcriptomic, and proteomic data of two different *Brassica napus* cultivars with either high or low oil content. The objective of the authors was to investigate the complex interplay between regulatory SNPs (rSNPs), transcription factors (TFs) and their corresponding genes. The study provides the first genome-wide collection of rSNPs for *B. napus*, among which those significantly associated with the phenotype were identified with a machine learning approach.

For the assessment of the importance of single rSNPs, regarding their association to the *B. napus* cultivars, the authors implemented a machine learning approach known as random forest (RF)-based feature selection (i.e. the Boruta algorithm), which calculates the importance of a variable in a dataset by utilizing multiple decision trees. Using the Boruta algorithm, they identified a total of 1,141 genome-wide regulatory SNPs significantly associated with oil content, a higher number of significant SNPs than conventional methods such as GWAS. By using a machine learning approach, they were thus able to identify and investigate the interplay between several rSNPs and TFs to unravel the regulatory mechanisms underlying several candidate genes of oil content in B. *napus*.

The authors emphasize the applicability of ML to handle various data types, such as multi-omics data, to find important biological features influencing a trait of interest.

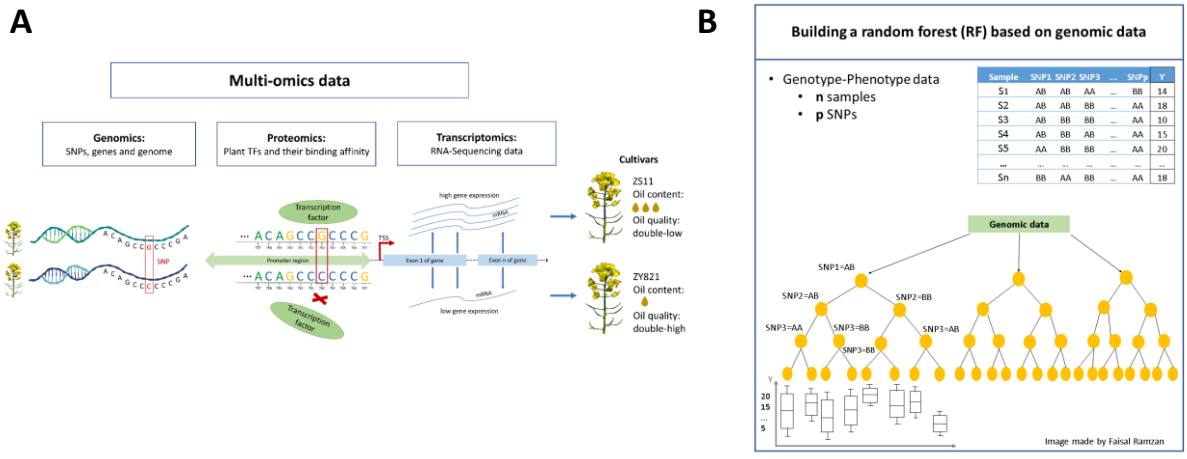

Figure 1. *The basic scheme of the study illustrates that different types of multi-omics data were jointly analyzed to investigate rapeseed cultivars with high or low oil content (A).*
With Random Forest, the variable importance is calculated using genomics data by constructing several decision trees based on random subsets of data. At each tree split, the sample data are divided into subgroups based on the best splitting variable (i.e. SNP) in order to obtain the best prediction gain (B).

Researcher: Torsten Pook
Paper: Using local convolutional neural networks for genomic prediction
Link: https://www.frontiersin.org/articles/10.3389/fgene.2020.561497/full
The methods commonly applied in the prediction of breeding values are based on linear mixed models or Bayesian approaches. A similarity between these approaches is that they are mainly based on modeling additive genetic effects. One challenge of these methods that the authors of this paper outlined is the computational demand when scaling to large datasets or including additional input factors such as weather, soil, or housing conditions into models. In their publication, Torsten and CiBreed colleagues proposed a method referred to as Local Convolutional Layers (LCL) which is an extension to a class of artificial neural networks (ANNs) known as Convolutional Layers. In their analysis with simulated and real datasets, they found that that the application of local convolutional neural networks (LCNN) substantially improves the accuracy of genomic prediction relative to more frequently applied ANNs architectures like multi-layer perceptrons (MLP) and classical convolutional neural networks (CNN). They found that the predictive ability of state-of-the-art methods like GBLUP outperformed the proposed LCNN method in small datasets. However, LCNN outperformed GBLUP in large datasets.

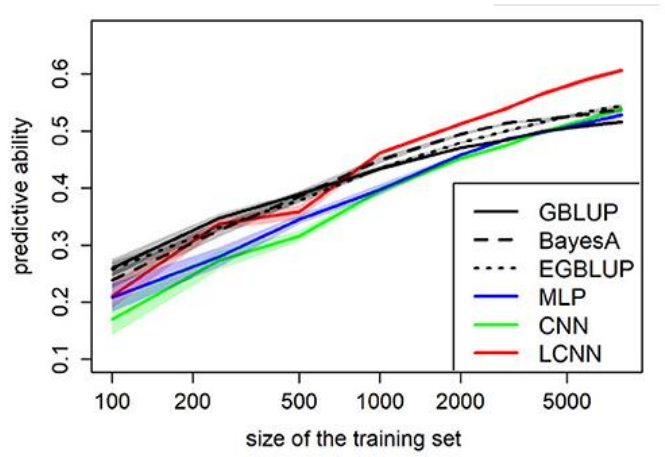

Figure 2. Performance of different prediction methods according to the size of the training set. The advantage of using local convolutional neural networks becomes substantial at larger training set sizes (>2000).

Researcher: Cathy Westhues
Paper: Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks
Link: https://www.frontiersin.org/articles/10.3389/fpls.2021.699589/full

In their article, Cathy Westhues et al. investigated the use of gradient boosted decision trees (GBDT) algorithms for prediction of quantitative traits (grain yield and plant height) in maize with genomic and environmental predictor variables. One of the objectives of the study was to determine whether these non-linear prediction methods were competitive with linear mixed models across four cross-validation (CV) schemes relevant in plant breeding, and to examine the effect of some environmental predictors on grain yield. One advantage expected by these sophisticated models is improved modeling of nonlinear interactions between

genomic-derived predictors (here, principal components from SNPs data) and environmental factors.

The authors found a slight gain in using GBDT with weather data in some challenging prediction scenarios, such as predicting genotypes in a new year for grain yield. Variable importance scores from the full model dataset were obtained to examine the relative contribution of different covariates in the model, showing the importance of abiotic stress factors such as high temperatures at flowering time and precipitation at different growth stages. However, authors emphasize that GBDT frameworks might yield even more significant improvements of predictive ability by including a larger amount of phenotypic data, necessary to capture genotype-by-environment interactions. Regularization is also an important component of ML applications, for instance the authors found that three hyperparameters had a significant influence on model performance. Prior feature engineering aiming at reducing data dimensionality and at tailoring relevant input predictors in the context of plant breeding - for instance by computing stress covariates from original weather data – was also critical and could potentially be further investigated in future studies.

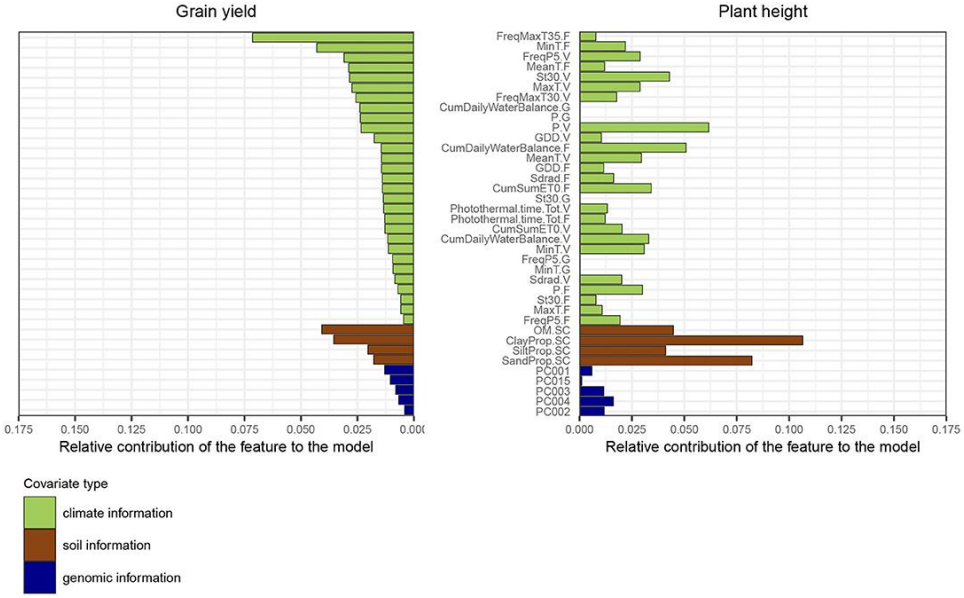

***Figure 3**. Feature importance ranking based on the average relative gain per feature obtained with the model XGBoost using environmental and genomic covariates, for the two traits grain yield and plant height, obtained from multi-environment hybrid maize field datasets. The gain represents the improvement in accuracy when using a feature for splitting, across all trees in the model. The order of features is based on feature performance within covariate class for the trait grain yield.*

Researcher: Martin Wutke
Paper: Detecting Animal Contacts—A Deep Learning-Based Pig Detection and Tracking Approach for the Quantification of Social Contacts
Link: https://doi.org/10.3390/s21227512

Another very interesting application of machine learning in animal breeding is phenotyping technologies. The scale and size of modern breeding farms make monitoring and caring for the health and welfare of animals by visual observation difficult. The monitoring of animal

social interactions is important to prevent stress-induced behaviours, which are often negatively correlated with the health, productivity, and welfare of animals.

Martin and colleagues applied a novel object detection method introduced for the detection of different body parts of pigs, which were used the detected points for tracking individual animals. Based on the tracking results, they identified specific animal contacts and automatically construct a social network highlighting distinct contact types. The machine learning approach overcomes the limitation of manual data collection in Social Network Analysis (SNA). An ML-based technique particularly provides improvements to current SNA methods by providing automated social network construction and increasing the power of understanding and identifying social interactions. The study showed the suitability of CNN-based object detection methods for the identification of distinct contact types and SNA construction. They evaluated the detection and tracking performance of the framework by utilizing well established evaluation metrics, and achieved state of the art performance results. They continue to develop this framework by expanding the training data and increasing the generalizability of the framework to a broader range of livestock environments.

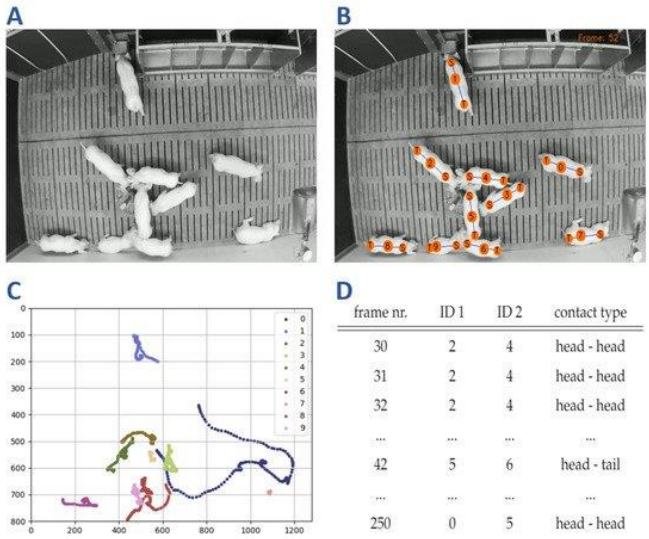

***Figure 4**. The social contact identification starts with a raw video frame (A). After detecting the shoulder and tail position (B), the trajectories are computed and analyzed over time (C). By identifying cases of close distances, a table of social contacts is constructed automatically (D).*

Machine learning algorithms, revolutionizing tools in plant and animal breeding?

Machine Learning presents great potential for plant and animal breeding. ML algorithms can improve the prediction of genomic prediction, particularly when part of the genetic variance is non additive (Azodi et al., 2019; Pérez-Enciso and Zingaretti, 2019). It is worth emphasizing that the applications of machine learning are not limited to the genomic aspects of breeding. Machine learning increases the ability of breeders to collect and harness primary and secondary phenotypes from plants and animals. Machine learning also increases the option of indirect data sharing amongst companies through technologies such as transfer learning.

However, despite their immense versatility and their ability to uncover complex relationships in datasets, ML algorithms should not be considered as a “magical approach” to all problems in breeding. For many phenotypic traits, ML does not yet consistently outperform parametric models such as GBLUP for genomic prediction (Abdollahi-Arpanahi et al., 2020; Bellot et al., 2018; Zingaretti et al., 2020). A common challenge faced by many applications of ML is the need for large training sets, so that these advanced methods are really able to capture hidden relationships in the dataset and outperform classical statistical approaches. This remains challenging, as phenotyping has now become the major bottleneck, while high-throughput genotyping techniques have enabled the generation of genotypic data at reduced costs. Finding the best model architecture or the best hyperparameters, and efficient handling of large high-dimensional datasets, are also important requirements for an efficient use of ML methods. Inadequate model regularization can notably contribute to overfitting issues, i.e. preventing the model to generalize well when confronted with new data. Besides these aspects, ML models are still often considered as black boxes, although various interpretation methods, such as feature importance (Fisher et al., 2019), partial dependence plots, accumulated local effects plots (Apley et al., 2020), saliency maps and bias mitigation techniques have progressively gained interest, thereby demonstrating that ML applications are not restricted to make highly accurate quantitative predictions or classifications (Molnar, 2020). Further studies are undoubtedly required for further improvement of ML methodologies, thereby enabling better selection decisions in plant and animal breeding.

Bibliography:

Thomas M. Mitchell. 1997. Machine Learning (1st. ed.). McGraw-Hill, Inc., USA.

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., & Shiu, S. H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics, 9(11), 3691-3702. https://doi.org/10.1534/g3.119.400498

Pérez-Enciso, Miguel, and Laura M. Zingaretti. "A guide on deep learning for complex trait genomic prediction." Genes 10.7 (2019): 553. https://doi.org/10.3390/genes10070553

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., Whitaker V. M. & Pérez-Enciso, M. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science*, *11*, 25.

Abdollahi-Arpanahi, R., Gianola, D. & Peñagaricano, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol **52,** 12 (2020). https://doi.org/10.1186/s12711-020-00531-z

Pau Bellot, Gustavo de los Campos, Miguel Pérez-Enciso, Can Deep Learning Improve Genomic Prediction of Complex Human Traits?, *Genetics*, Volume 210, Issue 3, 1 November 2018, Pages 809–819, https://doi.org/10.1534/genetics.118.301298

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 82(4), 1059-1086. https://doi.org/10.1111/rssb.12377

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.*, *20*(177), 1-81.