Revista Brasileira de Milho e Sorgo

Brazilian Journal of Maize and Sorghum

ISSN 1980 - 6477

Journal homepage: www.abms.org.br/site/paginas

MACHINE LEARNING FOR CROP SCIENCE: APPLICATIONS AND PERSPECTIVES IN MAIZE BREEDING

Alexandre Hild Aono[1](✉), Ricardo José Gonzaga Pimenta[1], Felipe Roberto Francisco[1], Anete Pereira de Souza[1,2] and Ana Carolina Lorena[3]

[1] Molecular Biology and Genetic Engineering Center (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil.
E-mail:alexandre.aono@gmail.com,
ricardojgpimenta@gmail.com,
felipe.roberto.francisco@gmail.com,
anete@unicamp.br.

[2] Department of Plant Biology, Institute of Biology (IB), University of Campinas (UNICAMP), Campinas, SP, Brazil.
E-mail: anete@unicamp.br

[3]Technological Institute of Aeronautics, São José dos Campos, SP, Brazil
E-mail: aclorena@ita.br

✉ Corresponding author

Abstract – Machine learning (ML) has been a major driver in complex data analysis in recent decades, allowing the mining of large databases. ML techniques allow the creation of computational models for prediction, pattern extraction and recognition, considering the premise that the computer acquires learning skills to perform a given task without being explicitly programmed for such a purpose. Driven by the efficiency of these techniques, several studies have demonstrated their wide range of applications and high potential for maize breeding. From the prediction of genetic values by omic data to applications of high-throughput phenotyping data, ML models have promoted advances in the species comprehension and assisted in the development of more effective tools for its breeding, driving expressive yield gains. In this context, this work presents the main contributions of ML in maize breeding, providing a broad view of the main studies and methodological perspectives in the area.

Keywords: Artificial intelligence, deep learning, high-throughput phenotyping, omics-based prediction

APRENDIZADO DE MÁQUINA NA AGRICULTURA: APLICAÇÕES E PERSPECTIVAS NO MELHORAMENTO DE MILHO

Resumo - O aprendizado de máquina (AM) tem sido um impulsionador na análise de dados complexos nas últimas décadas, permitindo a mineração de amplos bancos de dados. Técnicas de AM permitem a criação de modelos computacionais para predição, extração e reconhecimento de padrões, considerando a premissa de que o computador adquire habilidades de aprendizado para realizar uma dada tarefa sem ser explicitamente programado para tal. Impulsionados pela eficiência de tais técnicas, diversos estudos têm demonstrado a ampla gama de aplicações e elevado potencial no melhoramento de milho. Desde a predição de valores genéticos por dados ômicos a aplicações de tecnologias para fenotipagem de alto desempenho, modelos de AM vêm promovendo avanços no conhecimento da espécie e auxiliando no desenvolvimento de ferramentas mais efetivas para seu melhoramento, impulsionando ganhos produtivos expressivos. Nesse contexto, neste trabalho são apresentadas as principais contribuições do AM no melhoramento de milho, fornecendo uma ampla visão dos principais estudos realizados e perspectivas metodológicas na área.

Palavras-chave: Inteligência artificial, aprendizagem profunda, fenotipagem de alto rendimento, predição baseada em ômicas

How to cite

AONO, A. H.; PIMENTA, R. J. G.; FRANCISCO, F. R.;SOUZA, A. P.; LORENA, A. C. Machine learning for crop science: applications and perspectives in maize breeding. **Revista Brasileira de Milho e Sorgo**, v. 21, e1257, 2022.

DOI: https://doi.org/10.18512/rbms2022vol21e1257

1. Contextualization

Since the Neolithic revolution, plant breeding has gone through different stages (Ramstein et al., 2018), which Wallace et al. (2018) divide into three main periods throughout history. "Breeding 1.0" was based on the selection of individuals empirically; "Breeding 2.0" was characterized by the use of statistical tools (Jiang et al., 2020) to support such selection, and "Breeding 3.0" was defined by the use of linear regression models to correlate molecular markers with quantitative trait loci (QTLs). We are currently undergoing a new stage, "Breeding 4.0", which aims to assess biases and high dimensionality caused by the large number of markers in order to safely verify the effects of each locus (Ramstein et al., 2018).

In recent years, the amount of genomic data generated has achieved unprecedented levels, enhancing larger genetic gains through molecular breeding strategies (Dwivedi et al., 2020). This was only possible with the development of molecular biology techniques and high-throughput genotyping (Prohens et al., 2011). In order to coordinate this data generation and provide means for deciphering complex genetic relationships, high-throughput phenotyping approaches have also emerged in crop science as an indispensable tool for predictive breeding, encompassing image analysis, robotics and remote-sensors (Kim et al., 2020).

Maize (*Zea mays* L.*)* is one of the most important crops in the world, being a source of food, fodder and industrial products in tropical and subtropical regions (Ranum et al., 2014). In 2019, global production of maize was estimated at 1,14 gigatonnes, yielding over 7 billion dollars in exports in Brazil alone (FAO, 2021a, 2021b). Due to this economical relevance, maize is probably the best-studied crop to date, with large amounts of resources available and for which breeding schemes are highly optimized. In the interest of the large amount of complex data that has been made available with the rapid advancement of phenotyping and genotyping technologies, machine learning (ML) algorithms have arisen as a promising tool in maize research and breeding.

In this context, this article aims to present a review of the development of ML-based strategies in maize breeding, encompassing strategies for associating phenotypic data, dealing with high-throughput phenotyping approaches and modelling genotype-phenotype associations. In order to present the main advances in ML breeding strategies, summarize the state-of-the-art into such methodologies and provide methodological perspectives, this review also includes important concepts from computer science and data analytics.

2. Machine Learning Concepts and Definitions

Based on the first studies with programmable devices, the theoretical foundation of Computer Science (CS) started in the 1950s (Haigh, 2014). From a wide range of industrial applications to personal use, the development of CS has boosted the economy in many ways and created several facilities for the storage, management and processing of data. Together with such advancements, artificial intelligence (AI) formally started in the late 1950s as a subfield of CS (Osama et al., 2015), providing more ambitious perspectives on systems' capabilities by introducing the idea of autonomous technologies. CS is a vast field, including concepts from mathematics and engineering; AI encompasses all this theory, further supplying a means of automatically solving problems based on the idea that systems can think and

DOI: https://doi.org/10.18512/rbms2022vol21e1257

act rationally.

With the unprecedented increase of data generation in the last decades, efficient data analysis techniques have been increasingly demanded. All the steps required for producing relevant information from large datasets are part of a process known as knowledge discovery in databases (KDD), which encompasses data cleaning, integration, selection, transformation, mining and evaluation (Han et al., 2011). Turning data into knowledge requires data mining procedures, which are mainly accomplished by ML models; ML is a part of AI focusing on automating the process of knowledge extraction from data, which is usually performed by the induction of models able to identify patterns on data (Faceli et al., 2021). For such, ML techniques make use of programming techniques from CS and AI allied to methods from statistics, optimization and information theory, adapted for complex datasets and suitable for current hardware systems (Tarca et al., 2007).

Formalizing, let *X* be a data matrix containing *n* observations, which are described by *p* qualitative or quantitative variables or attributes. The input data may be either accompanied or not by a response variable *Y*, containing labels for each one of the *n* observations. These labels can be either quantitative, when a regression problem is configured, or qualitative, when they are named classes and a classification problem is configured instead. Techniques aiming at estimating a model for predicting a response variable based on information that rely only in the set *X* are considered part of supervised learning (SL). Unlabeled datasets are dealt with as an unsupervised learning (UL) task, which has a more descriptive nature, identifying patterns intrinsic to the dataset which can indicate the presence of clusters or associations between variables (Van Dijk et al., 2021).

Several mathematical formulations support the theory underlying ML methods. Currently, one of the most popular ML methods is deep learning (DL), a concept introduced in 2006 (Vargas et al., 2017). The development of deep neural networks has dramatically improved the predictive scenario for several important applications, including image recognition and genomics (Lecun et al., 2015). There are DL models for both SL and UL scenarios, incorporating an already established theory into high dimensional networks. Such methods have gained important visibility, mainly for supplying alternatives for processing large image databases, such as those present in the Imagenet dataset (Deng et al., 2009).

In a common SL setting, the definition of the most appropriate method for a predictive task relies on the evaluation of the produced models by validation strategies, i.e., defining random partitions for the dataset and creating the training and test sets accordingly (Figure 1). The most common data partitioning strategies are (Faceli et al., 2021): (i) hold-out, where the dataset is split into a percentage for training and a left-out percentage for testing; (ii) random subsampling where (i) is repeated *k* times with different random seeds for a most reliable evaluation; (iii) k-fold cross-validation with the definition of *k* mutually exclusive subsets of data, which are used to fit a model *k* times considering *k-1* folds for training and 1 left-out fold for testing; (iv) leave-one-out, an extreme case of cross-validation where *k*=*n*, the model is trained *n* times and at the *i*-th iteration, the model uses *n-1* observations for training and predicts the label of the *i*-th left-out observation; and (v) bootstrap, which samples data for training multiple times with replacement (Han et al., 2011).

Selecting the most appropriate model is not simple and depends on the complexity and size of the dataset – which must be compatible with the strengths and weaknesses of a given algorithm too (Olson et al., 2017). As discrepancies on predictions are expected for different ML workflows, it is common to test a wide range of techniques, contrasting the models' predictions to the real labels of the test data using metrics such as accuracy, precision, recall and F-measures, confidence intervals, comparative statistics, and receiver operating characteristic (ROC) curves (Han et al., 2011). There are several strategies for building predictive models (Bishop, 2006), including: (i) linear models (e.g. discriminant analysis, probabilistic generative models, logistic regression, and Bayesian approaches); (ii) neural networks and Bayesian neural networks; (iii) kernel based methods (e.g. support vector machine (SVM) and Gaussian processes); (iv) graphical models (e.g. Bayesian networks, decision trees, and random forests (RFs)); and (v) ensemble strategies as bagging, boosting (e.g. AdaBoost, and RFs). All of these methods aim at establishing a predictive model and have distinct biases regarding how such models are represented and searched (Figure 1).

For every ML algorithm, a related theoretical background for defining the predictive model exists. Therefore, to take advantage of these techniques and use them properly, it is necessary to understand the inherent model assumptions, functioning and capabilities, in order to define the model hyperparameters and also suitable data manipulations. Another important point regards the exclusion of part of the input variables in X when predicting Y. This process, known as feature selection (FS), is quite important for applications with high dimensional data, where there are often redundant, noisy and irrelevant features. Through different evaluation criteria, FS methods aim at defining an optimal feature subset which can supply a more effective input for ML model creation (Kumar and Minz, 2014).

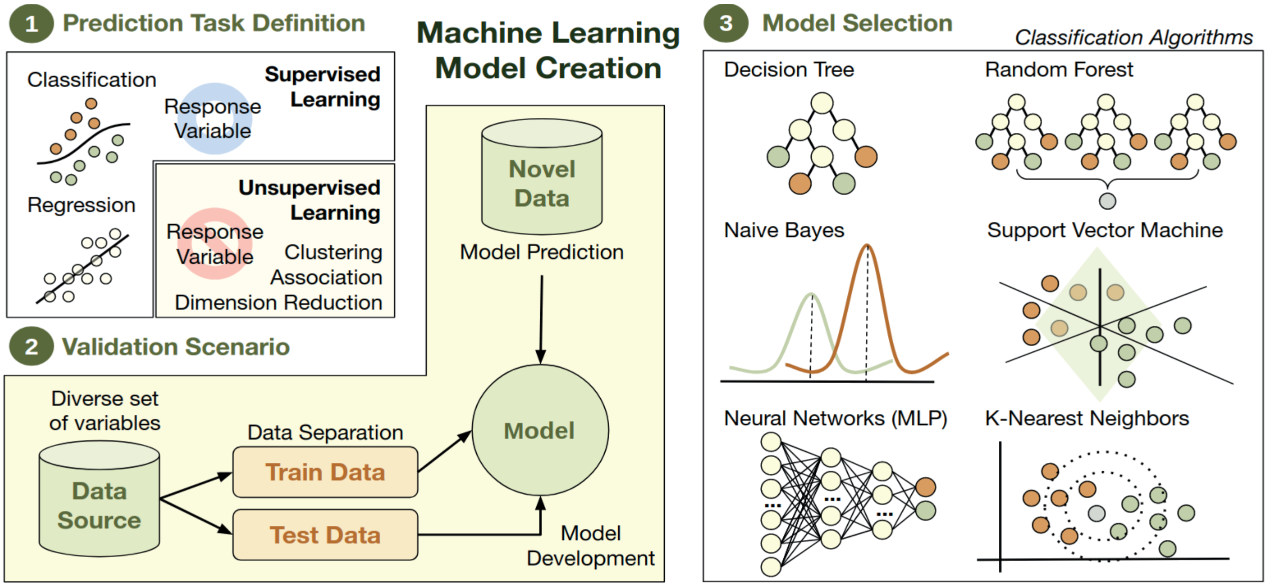

Figure 1. Machine learning (ML) workflow and main ML models for predictive tasks.

DOI: https://doi.org/10.18512/rbms2022vol21e1257

3. Maize Breeding and Machine Learning

3.1. Machine Learning Systems in Maize

The results of a search including a few keywords (maize, corn, crop, agriculture, machine learning and artificial intelligence) on the Scopus and Web of Science databases using the bibliometrix R package (Aria and Cuccurullo, 2017) are presented in Figure 2. This inspection highlights the rapid advancement in the use of ML techniques in agricultural research; while works in this area have been published since the early 2000s, an explosive trend is observed in the last two years. Additionally, a large proportion of the publications that involve ML and crops also involve maize, indicating that this crop is the subject of study of much of the research in this field. The analysis also shows that two countries with most publications employing ML in maize research are also the world's largest producers of this crop – USA and China; in Brazil, the third largest maize producer (FAO, 2021a), studies in this field are not as abundant, but still noteworthy considering the global scenario.

One of the most basic applications of ML in agriculture is to predict performance based on large phenotypic and environmental datasets, which has great potential to assist in the selection and recommendation of locally adapted cultivars, as well as in the optimization of agricultural practices. Such an approach was first taken for maize by Kaul et al. (2005), who used artificial neural networks to predict yield in Maryland (USA) based on historical data of yield and climate; these authors found that the ML models resulted in consistently more accurate predictions than linear regression. Many

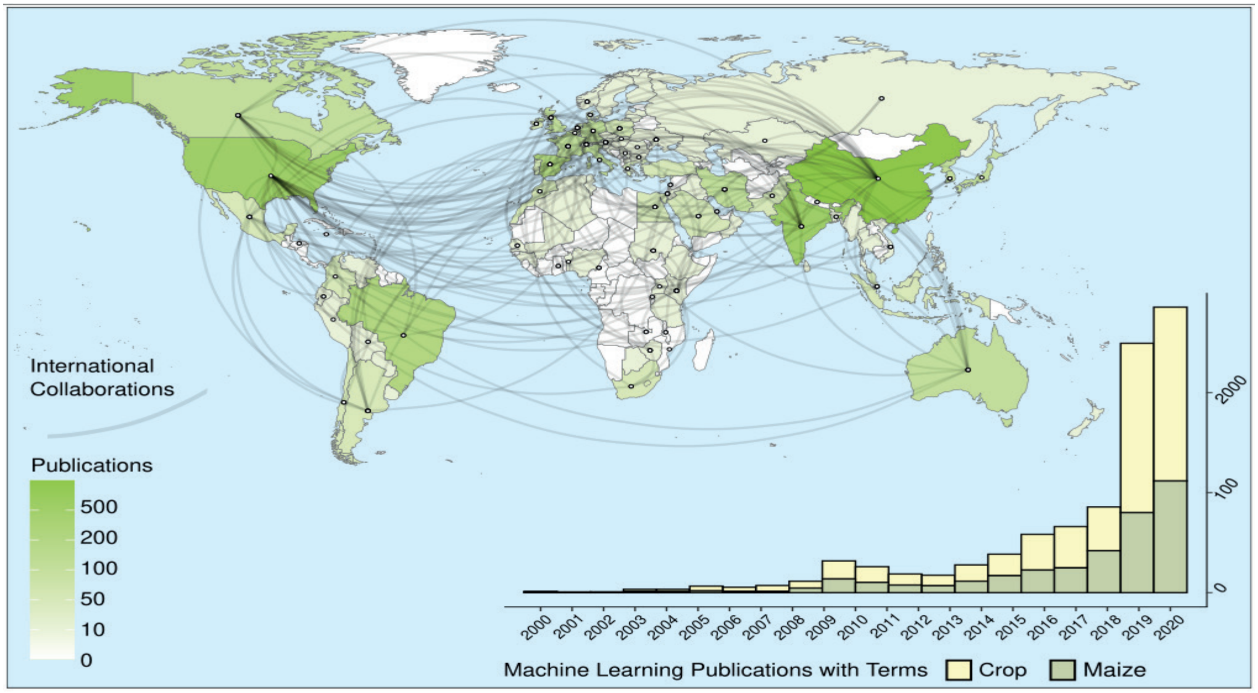

Figure 2. Growth of studies using machine learning in crop and maize research and distribution of publications per country.

other studies have employed various ML strategies to forecast maize yield based on genotype, management and climatological information, often achieving lower error rates with ML methods than with linear regression techniques (Folberth et al., 2020; Chen et al., 2021; Correndo et al., 2021; Prasad et al., 2021).

This type of predictive model is set to be especially relevant in the current scenario of climate change (Wheeler and Von Braun, 2013). Maize is already being severely affected by these changes, as clearly experienced in the 2012 USA harvest (Rippey, 2015). Thus, understanding the impact of climate change on the yield crop is a key step to enhance its resilience and prevent further losses. Considering the realistic scenario of a 2°C increase in temperature by 2050, Leng and Hall (2020) used ML algorithms to accurately forecast a decrease in maize yield of 13.5% by 2050. These results, which are backed-up by similar studies employing ML (Fan et al, 2020), evidentiate and give cues to the need of a directed effort of development of drought- and heat-tolerant maize cultivars by breeders, as well as of the advancement of agricultural techniques.

Yet, the pivotal role of genotype in the yield of any crop is undisputable. Since the late 1960s, maize yields have drastically increased with the development of heterozygous hybrids from single crosses of inbred lines, which outperform both parents as a consequence of heterosis (Crow et al., 1998). Thus, predicting the performance of these hybrids is historically one of the most targeted, but also challenging, tasks in maize breeding. While several traditional approaches based on frequentist statistics were first employed for this task (Bernardo, 1996), ML approaches are recently being used with the same objective (Khaki et al., 2020; Sarijaloo et al., 2021).

3.2. Plant Phenotyping

Several measures can be useful for phenotyping an individual according to its potential in the field, providing an extragenic characterization at different scales (Yang et al., 2020). Although such a process has been automated, phenotyping thousands of plants is still challenging for breeders (Araus et al., 2018). In this sense, the idea of phenomics has been introduced in crop breeding, encompassing the study of high-throughput phenotypic data acquired through multispectral, hyperspectral, fluorescence, and thermal sensors and imagers, which are then processed through high-level information technologies (Araus et al., 2018; Yang et al., 2020).

Following the increase in the interest in high-throughput phenotyping, ML approaches have become popular in phenomics. The dramatic boom in data generation coupled with the elevated potential of preventing yield losses, identifying desirable phenotypes in large fields, and mostly obtaining increases in performance through artificial selection, have turned the breeding bottleneck to data management and processing (Shakoor et al., 2017). For maize, the most desirable traits, including yield, quality, flowering time, and stress resistance (Jiang et al., 2020), have already been phenotyped through remote sensing and the related images processed through image processing methods coupled with ML techniques.

Differently from structured datasets, images have no intrinsic features to be used for prediction. Although the image pixels contain all the information needed for prediction, such information needs to be structured, and common approaches focus on characterizing an image using descriptors. For plant phenomics, such characterization is based on vegetation indices (VIs) (Vergara-Díaz et al., 2016). In maize, VIs have been used for characterizing several types of images and predicting important phenotypes through ML, such as yield and biomass (Jeffries et al., 2019; Zhang et al., 2020).

Although very popular in plant breeding, VIs do not capture the entire information present on the images (Araus et al., 2018); in some scenarios, this lack of features can hinder the construction of effective models. As an alternative, DL approaches with convolutional operations have been suggested as the most suitable solution. A convolutional neural network (CNN) is a type of neural network with several matrix transformations, enabling the acquisition of abstract features for a given image. Although the concept of DL is related to networks with a large number of layers (Lecun et al., 2015), one of the most popular DL networks are CNNs, which have surpassed the classical methods performance for pattern recognition in several areas (Rawat and Wang, 2017). CNN kernel filters enable these models to automatically learn the domain features instead of relying on specific image descriptors (Aloysius and Geetha, 2017).

For maize breeding, DL has shown promising results for predictive tasks on images, including maize segmentation and detection (Liu et al., 2020a), kernel and tassel identification and evaluation (Liu et al., 2020b; Zhang et al., 2020; Khaki et al., 2021), seed analyses (Huang et al., 2019), and disease/ stress prediction (Condori et al., 2017; Jiang et al., 2019; Sibiya and Sumbwanyambe, 2021). In addition to providing ways of measuring maize yield, these methods have already been shown to be efficient in predicting this trait (Jin et al., 2020; Khaki et al., 2021).

As the network architecture definition impacts CNN performance, different approaches have been tested in maize image analyses, including the creation of novel architectures (Jin et al., 2020; Zhang et al., 2020), the use of previously described combinations with modifications (Jiang et al., 2019; Sun et al., 2020), and also attempts on using pre-trained architectures through transfer learning (Condori et al., 2017; Huang et al., 2019; Liu et al., 2020b; Sibiya and Sumbwanyambe, 2021). Such pre-trained networks are usually developed using the ImageNet dataset (Deng et al., 2009), which is composed of ca. 14 million images from several resources, organized into over 20,000 categories. Several networks have been developed for predicting such labels, including VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and GoogLeNet (Szegedy et al., 2014), all of which have shown promising results for maize image analyses (Condori et al., 2017; Huang et al., 2019; Jiang et al., 2019; Liu et al., 2020a; Sun et al., 2020; Sibiya and Sumbwanyambe, 2021). For non-expert users, the implementation of such ML structures might be challenging. In this sense, DL tools exclusive for maize have also been developed (Baweja et al., 2018; Shete et al., 2020; Barman et al., 2021; Zhou et al., 2021).

3.3. Genomic-based Prediction

One of the main obstacles faced by plant breeders is the stagnation in genetic gains (Cooper et al., 2020). Over the last few years, the insertion of genomics into breeding pipelines has enabled great reductions in the duration of cycles (Voss-Fels et al., 2019). Through marker-assisted selection (MAS) techniques, the incorporation of predictive models based on genomic information to assess the potential of new varieties has brought considerable gains in maize breeding (Riedelsheimer et al., 2012). Furthermore, with the employment of genomic selection (GS) models, important traits such as productivity and resistance to various stresses could be incorporated into MAS (Riedelsheimer et al., 2012).

Based on the development of a predictive model for a training population containing both genotypic and phenotypic information, GS allows the evaluation of the field potential of new genotypes for which only genotypic information needs to be obtained, reducing costs associated with future phenotypic field evaluations (Jannink et al., 2010). Although having been first proposed in 2001 (Meuwissen et al., 2001), GS only became popular with the cheapening of sequencing technologies. It was first employed in maize breeding in 2007 (Bernardo and Yu, 2007) and has since been demonstrated to be a valuable tool in this area (Riedelsheimer et al., 2012; Môro et al., 2019).

ML approaches have been explored as an alternative to traditional GS models, which are mainly based on the use of genomic best linear unbiased prediction (GBLUP) methods and their variations, eventually including genotype-environment interactions (Lado et al., 2016). Among the most frequently tested algorithms we can mention those based on neural networks, which have led to the hypothesis that DL tools could further leverage the predictive scenario of GS. Due to the power of DL for prediction, the usage of such strategy for the definition of genomic selection models has been widely tested, however with controversial results, which vary depending on the species and the trait, and also on the DL architecture and the hyperparameter optimization methodology employed (Crossa et al., 2019). In maize breeding, the overall attempts of employing DL strategies have been successful or presented equiparable performances to traditional methodologies, depending on the statistical methods used, the trait genetic characteristics and the presence of genotype/environment interactions. The most used DL strategies have been based on feed-forward multilayer perceptron (MLP) architectures (Heslot et al., 2012; González-Camacho et al., 2016), however with few attempts on other structures, such as deep belief networks (Rachmatia et al., 2017), Bayesian regularized networks (Pérez-Rodríguez et al., 2020), and CNNs (Azodi et al., 2019; Pook et al., 2020).

Other ML approaches tested for maize prediction are based on SVM and RF algorithms (Heslot et al., 2012; Qiu et al., 2016; Azodi et al., 2019; Badji et al., 2020; Li et al., 2020), but in general the results were equivalent or worse than GBLUP based models. Grain yield and moisture and flowering time are the traits most frequently targeted by these studies (González-Camacho et al., 2016; Heslot et al., 2012; Azodi et al., 2019; Li et al., 2020), and have sometimes been evaluated under the increasingly relevant condition of drought stress (Qiu et al., 2016; Rachmatia et al., 2017). Other characteristics addressed include plant and ear height (Azodi et al., 2019; Li et al., 2020), resistance to pests (Badji et al., 2020) and diseases (Pérez-Rodríguez et al., 2020), ear leaf traits, stalk strength (Li et al., 2020), and grain color and starch content (Yin et al., 2020).

One of the main difficulties faced by researchers in creating a predictive model for GS lies in the high dimensionality caused by the high number of markers in relation to the number of genotypes available (Ramstein et al., 2018). Most popular GS models can handle the number of available markers, but alternatives to further increase the predictive model capability have been developed. In traditional statistical approaches, an alternative has been the introduction of biological information in the model, prioritizing certain genomic regions. The incorporation of QTL information obtained through linkage mapping and genome-wide association studies, for instance, has brought beneficial results with the inclusion of markers as fixed effects (Rice and Lipka, 2019; Liu et al., 2020a,b).

The inclusion of such markers on GS models was first tested by Bernardo (2014), who suggested that the inclusion of QTLs can be occasionally beneficial. For maize, there have been some controversial results in inserting such regions as fixed effects. Although Sousa et al. (2019) observed accuracy improvements, Rice and Lipka (2019) found lower accuracies together with an increase in the bias of the predicted breeding values, further corroborated by Galli et al. (2020). Another tested approach for dealing with such high dimensionality and improving the model predictions is the selection of a subgroup of markers for creating the GS models, which also reduces genotyping costs. In addition to statistical models for prioritizing a subset of markers to be used for prediction (Qiu et al., 2016), ML-based approaches have also been incorporated into this subset definition in maize (Ramstein et al., 2018) and in other crops as well (Li et al., 2018; Aono et al., 2020), showing promising results.

Another challenge in maize breeding that can be tackled by genomic-based approaches are clustering analyses, which have fundamental importance to evaluate population genetic structure. These analyses are useful for breeding in many aspects, being employed in diversity studies and also as a factor for genome-wide association mapping models (Pressoir and Berthaud, 2004a,b; Vigoroux et al., 2008). Various methods can be used for such assessments, some of which involve ML. López-Cortez et al. (2020), for instance, showed that a combination of hierarchical clustering with a deep autoencoder-based data preprocessing step was the most effective method to assign maize inbred lines to clusters.

Furtherly, in outcrossing crops such as maize, clustering can be used to classify parental lines into heterotic groups that, when inter-crossed, produce hybrids with superior performance (Melchinger and Gumber, 1998). This classification for parental choice has long been performed for maize, initially with low-throughput molecular markers and traditional distance-based clustering methods (Dias et al., 2004). Ornella and Tapia (2010) employed ML-based methods for this task using groups with diverse heterotic patterns and microsatellite-derived attributes; depending on the dataset structure, SVMs performed equally well or better than Bayesian and simple logistic functions, representing a valuable strategy.

4. Perspectives

One of the main concerns of both plant and animal breeding programs is how to reach elevated genetic gains in a world scenario with climate instability, food insecurity and exponential population growth. In addition to smart breeding designs (Singh et al., 2020), systems with minimal environmental impacts are required (Yu and Li, 2021); this can only be achieved with increases in yield production from the existing land through novel and efficient breeding techniques (Meng et al., 2018). In addition to the wise introduction of genetic diversity in maize breeding programs (Swarup et al., 2021), the most significant improvements from the last few years in modern breeding have been due to high-throughput data generation strategies, mostly from genomics and phenomics.

Coupling high-quality phenotypes with

a vast number of molecular markers widely spread throughout the genome enabled deeper characterizations of complex traits; however, the complete phenotypic definition is caused by a broader cascade of molecular mechanisms, including transcripts, proteins, metabolites, and their interactions (Muthamilarasan et al., 2019). High-precision phenotyping methods (Yang et al., 2020), feasible omics data acquisition (Pazhamala et al., 2021), robust AI based methodological analyses (Beans, 2020) and genomic editing (Mackelprang and Lemaux, 2020) are now being proposed as an integrated way of assisting breeding (Figure 3).

The theoretical background of data analytics has been gradually incorporated into maize breeding best practices, and AI techniques have provided a different perspective on how to create breeding predictive models; however, little is known on what omic layers should be combined for effective predictions and what approaches are required for that. In maize, there are few attempts on including other omics sources than genomics in complex trait predictive models, with positive results on increasing prediction accuracies (Guo et al., 2016; Kremling et al., 2019). Through statistical linear mixed effect model methods, gene expression (Guo et al., 2016; Kremling et al., 2019) and metabolite quantifications (Guo et al., 2016) were coupled with genetic markers and could increase the model accuracies through the capture of more effects explaining the phenotypic variation. In this sense, multi-omics ML predictions have been indicated as a promising tool for breeding (Tong and Nikoloski, 2021).

There is a great potential of smart strategies for reducing the marker data for the future of maize GS (Washburn et al., 2020). This subset definition has been indicated in the previous section as an effective tool for increasing prediction accuracies; however, there is still a lack of appropriate methodologies for this objective, especially based on multi-omics

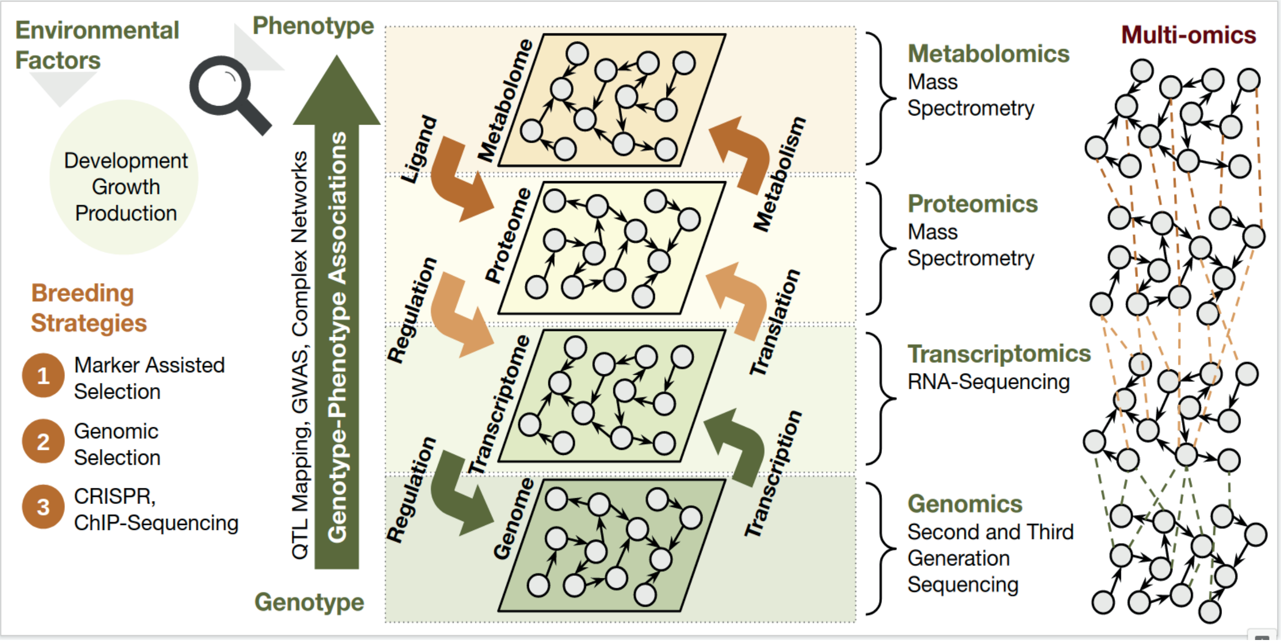

Figure 3. Multi-omics approaches.

integration. A common approach for elucidating such causal loci is by using gene annotations and marker associations (Dai et al., 2020), supplying indicatives for regulatory regions and molecular associations (Mejía-Guerra and Buckler, 2019). More initiatives on unraveling maize gene functions through ML are expected, such as the inclusion of explainable AI methods (Rai, 2020; Linardatos et al., 2021), supplying means of deciphering the "black box" created by most of these algorithms.

Maize breeding is moving towards a rational design of crops, including efforts for characterizing the plant architecture conferring a desirable trait. In addition to ML methods, systems biology techniques have also been incorporated into maize studies (Zhou et al., 2020), simulating complex systems through network approaches. Despite supplying powerful methodologies for assisting the prioritization of causal genes (Schaefer et al., 2018), such integrated methods can also provide hints on inter-omics associations, which is expected to be included in further maize prediction systems.

Finally, ML can also be of assistance in another gear of Breeding 4.0: gene editing. The application of this group of technologies has quickly advanced with the development and popularization of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) systems. This gene editing system has two key components: (1) a customized guide RNA (gRNA) which directs (2) a Cas9 endonuclease to the target site to be edited in the genome (Doudna and Charpentier, 2014). CRISPR/Cas9 has an extremely wide range of applications and is set to be extensively employed in plant biotechnology, as many countries are framing gene-edited crops under less strict legal regulations than first generation genetically modified organisms (Gupta et al., 2021).

In addition to the contributions to identify causal loci of phenotypes of interest that can be targeted by gene editing, ML can also be used in several steps of the design of CRISPR/Cas9 systems. Current models and softwares can be used to: identify high-efficiency target sites for editing (O'brien et al., 2019); predict gRNA on-target efficiency (Abadi et al., 2017; Wang et al., 2020) and off-target activity (Vinodkumar et al., 2021) and to optimize their design (Xue et al., 2018); and identify potential anti-CRISPR proteins (Eitzinger et al., 2020). ML has too been employed to identify the outcomes of gene editing, being used to detect RNA editing sites (Xiong et al., 2017) and to discriminate CRISPR/Cas9-induced mutant rice seeds using imaging (Feng et al., 2017). Various CRISPR/Cas systems have been developed for maize in the last few years (Svitashev et al., 2015; Qi et al., 2016), but this crop is yet to profit from these numerous associated ML tools. While there still much room for improvements in these techniques, they are advancing rapidly and already allow the algorithmical design of CRISPR experiments, which should further optimize gene editing efficiency from an experimental and economical point of view (O'brien et al., 2021).

As demonstrated throughout this review, ML approaches are an efficient tool for maize breeding, but the creation of efficient ML systems depends on CS expertise, which might transcend breeders' knowledge. In this sense, automated ML initiatives offer the possibility for non-specialists to easily create end-to-end ML pipelines (He et al., 2021). The idea is the estimation of a model from one execution, without the need of setting hyperparameters, selecting features, preprocessing data and generating the model (Weng, 2019). Although this is a recent topic in CS, there are already libraries destined for Automated

ML (AutoML) (Zimmer et al., 2021), with a growing number of researchers trying to increase the efficiency of the existing approaches as well (He et al., 2021). It is expected that, for the next few years, AutoML approaches will facilitate the use and applicability of AI based systems in maize breeding.

5. Final Remarks

AI is a vast field of CS, encompassing highly efficient algorithms for data analysis, structured in ML methods. For maize breeding, such initiatives have already been employed in diverse applications, showing promising results with virtually endless opportunities that respond to the current scientific scenario of diverse and massive data generation. From phenotypic evaluations to integrated data predictive systems, this technological reservoir of methodologies is required for advancing to the Breeding 4.0 era. With the advent and expansion of genome editing technologies, deeper molecular characterization of crops is required and, in respect to the current maize scenario, such advance is only possible with robust CS-based approaches.

Acknowledgements

This work was supported by fellowships from the São Paulo Research Foundation (2019/03232-6 to AHA, 2019/21682-9 to RJGP, and 2018/18985-7 to FRF), by a grant (Computational Biology Program - 88882.160095/2013-01 to APS) from the Coordination for the Improvement of Higher Education Personnel and by a research fellowship to APS (312777/2018) from the National Council for Scientific and Technological Development.

References

ABADI, S.; YAN, W. X.; AMAR, D.; MAYROSE, I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. **PLoS Computational Biology**, v. 13, n. 10, p. e1005807, 2017. DOI: https://doi.org/10.1371/journal.pcbi.1005807.

ALOYSIUS, N.; GEETHA, M. A review on deep convolutional neural networks. In: INTERNATIONAL CONFERENCE ON COMMUNICATION AND SIGNAL PROCESSING, 2017, Chennai, India. **Proceedings**… New York: IEEE, 2017. p. 588-592.

AONO, A. H.; COSTA, E. A.; RODY, H. V. S.; NAGARI, J. S., PIMENTA, R. J. G.; MACINI, M. C.; SANTOS, F. R. C.; PINTO, L. R.; LANDELL, M. G. D A.; SOUZA, A. P.; KUROSHU, R. M. Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. **Scientific Reports**, v. 10, n. 1, p. 1-16, 2020. DOI: https://doi.org/10.1038/s41598-020-77063-5.

ARAUS, O. J. L.; KEFAUVER, S. C.; ZAMAN A. M.; OSLEN, M. S.; CAIRNS, J. E. Translating high-throughput phenotyping into genetic gain. **Trends in Plant Science**, v. 23, n. 5, p. 451-466, 2018. DOI: https://doi.org/10.1016/j.tplants.2018.02.001

ARIA, M.; CUCCURULLO, C. bibliometrix: An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959-975, 2017. DOI: https://doi.org/10.1016/j.joi.2017.08.007.

AZODI, C. B.; BOLGER, E.; MCCARREN, A.; ROANTREE, M.; DE LOS CAMPOS, G.; SHIU, S. H. Benchmarking parametric and machine learning models for genomic prediction of complex traits. **G3: Genes, Genomes, Genetics**, v. 9, n. 11, p. 3691-3702, 2019. DOI: https://doi.org/10.1534/g3.119.400498.

BADJI, A. MACHIDA, L.; KWEMOI, D. B.; KUMI, F.; OKII, D.; MWILA, N.; RUBAIHAYO, P. Factors Influencing Genomic Prediction Accuracies of Tropical Maize Resistance to Fall Armyworm and Weevils. **Plants**, v. 10, n. 29, p. 1-22. 2020. DOI: https://doi.org/10.3390/plants10010029.

BARMAN, U.; SAHU, D.; BARMAN, G. G. A deep learning based android application to detect the leaf diseases of maize. In: INTERNATIONAL CONFERENCE ON MATHEMATICS AND COMPUTING, 6., 2021. **Proceedings**… Singapore: Springer, 2021. p. 275-286.

BAWEJA, H. S.; PARHAR, T.; MIRBOD, O.; NUSKE, S. Stalknet: a deep learning pipeline for high-throughput measurement of plant stalk count and stalk width. In: HUTTER, M.; SIEGWART, R. (ed.). **Field and service robotics.** Cham: Springer, 2018. p. 271-284.

BEANS, C. Inner Workings: Crop researchers harness artificial intelligence to breed crops for the changing climate. P**roceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 44, p. 27066-27069, 2020. DOI: https://doi.org/10.1073/pnas.2018732117.

BERNARDO, R. Best linear unbiased prediction of maize single-cross performance. **Crop Science**, v. 36, n. 1, p. 50-56, 1996. DOI: https://doi.org/10.2135/cropsci1996.0011183X003600010009x.

BERNARDO, R. Genome-wide selection when major genes are known. **Crop Science**, v. 54, n. 1, p. 68-75, 2014. DOI: https://doi.org/10.2135/cropsci2013.05.0315.

BERNARDO, R.; YU, J. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, v. 47, n. 3, p. 1082-1090, 2007. DOI: https://doi.org/10.2135/cropsci2006.11.0690.

BISHOP, C. M. **Pattern recognition and machine learning**. Berlin: Springer, 2006. 738 p.

CHEN, X.; FENG, L.; YAO, R.; WU, X.; SUN, J.; GONG, W. Prediction of Maize Yield at the City Level in China Using Multi-Source Data. **Remote Sensing**, v. 13, n. 1, p. 146, 2021. DOI: https://doi.org/10.3390/rs13010146

CONDORI, R. H. M.; ROMUALDO, L. M.; BRUNO, O. M.; CERQUEIRA, P. H. de. Comparison between traditional texture methods and deep learning descriptors for detection of nitrogen deficiency in maize crops. In: IEEE CONFERENCE ON COMPUTER VISION WHORKSHOP, 2017, Venice. **Proceedings**… New York: IEEE, 2017. p. 7-12.

CORRENDO, A. A.; ROTUANDO, J. L.; TREMBLAY, N.; ARCHONTOULIS, S., COULTER, J. A.; RUIZ-DIAZ, D.; FRANZEN, D.; FRAZLUEBBERS A. J.; SCHWALBERT, R.; WILLIAMS, J.; MESSINA; D. C.; CIAMPITTI, A. A. Assessing the uncertainty of maize yield without nitrogen fertilization. **Field Crops Research**, v. 260, p. 107985, 2021. DOI: https://doi.org/10.1016/j.fcr.2020.107985

COOPER, M.; TECHNOW, F.; MESSINA, C.; GHO, C.; TOTIR, L. R. Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. **Crop Science**, v. 56, n. 5, p. 2141-2156, 2016. DOI: https://doi.org/10.2135/cropsci2015.08.0512

COOPER, M.; TANG, T.; GHO, C.; HART, T.; HAMMER, G.; MESSINA, C. Integrating genetic gain and gap analysis to predict improvements in crop productivity. **Crop Science**, v. 60, n. 2, p. 582-604, 2020 DOI: https://doi.org/10.1002/csc2.20109

CROSSA, J.; MARTINI, J. W.; GIANOLA, D.; PÉREZ-RODRÍGUES, P.; JARQUIN, D.; JULIANA, P.; MONTESINO-LÓPEZ, O.; CUEVAS, J. Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. **Frontiers in Genetics**, v. 10, p. 1168, 2019. DOI: https://dx.doi.org/10.3389%2Ffgene.2019.01168

CROW, J. F. 90 years ago: the beginning of hybrid maize. **Genetics**, v. 148, n. 3, p. 923-928, 1998. DOI: https://doi.org/10.1093/genetics/148.3.923

DAI X.; XU, Z.; LIANG, Z.; TU, X.; ZHONG, S.; SCHNABLE, J. C.; LI, P. Non-homology-based prediction of gene functions in maize (*Zea mays* ssp. mays). **Plant Genome**, v. 13:e20015, p. 1-14, 2020. DOI: https://doi.org/10.1002/tpg2.20015

DENG, J.; DONG, W.; SOCHER, R.; LI, L. J.; LI, K.; FEI-FEI, L. Imagenet: a large-scale hierarchical image database. In: IEEE CONFERENCE ON COMPUTER VISION AND OATTERN RECOGNITION, 2009, Miami. **Proceedings**… New York: IEEE, 2009. p. 248-255.

DIAS, L. A. D. S.; PICOLI, E. A. D. T.; ROCHA, R. B.; ALFENAS, A. C. A priori choice of hybrid parents in plants. **Genetics and Molecular Research**, v. 3, n. 3, p. 356-368, 2004.

DOUDNA, J. A.; CHARPENTIER, E. The new frontier of genome engineering with CRISPR-Cas9. **Science**, v. 346, n. 6213, 2014. DOI: https://doi.org/10.1126/science.1258096

DWIVEDI, S. L.; GOLDMAN, L.; CECCARELLI, S; ORTIZ, R. Advanced analytics, phenomics and biotechnology approaches to enhance genetic gains in plant breeding. **Advances in Agronomy**, v. 162, p. 89-142, 2020. DOI: https://doi.org/10.1016/bs.agron.2020.02.002.

DREYFUS, S. The numerical solution of variational problems. **Journal of Mathematical Analysis and Applications**, v. 5, n. 1, p. 30-45, 1962. DOI: https://doi.org/10.1016/0022-247X(62)90004-5.

EITZINGER, S; ASIF, A.; WATTERS, K. E.; LAVARONE, A. T.; KNOTT, G. J.; DOUDNA, J. A.; MINHAS, F. U. A. A. Machine learning predicts new anti-CRISPR proteins. **Nucleic Acids Research**, v. 48, n. 9, p. 4698-4708, 2020. DOI: https://doi.org/10.1093/nar/gkaa219.

ERTIRO, B. T.; LABUSCHAGNE, M.; OSLEN, M.; DAS, B.; PRASANNA, B. M.; GOWDA, M. Genetic dissection of nitrogen use efficiency in tropical maize through genome-wide association and genomic prediction. **Frontiers in Plant Science**, v. 11, p. 474, 2020. DOI: https://doi.org/10.3389/fpls.2020.00474

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. C. P. L. F. (2021). **Inteligência Artificial:** uma abordagem de aprendizado de máquina, 2nd edition, *Rio de Janeiro: LTC*.

FAN, J; JINTRAWET, A.; SANGCHYOSWAT, C. The Relationships Between Extreme Precipitation and Rice and Maize Yields Using Machine Learning in Sichuan Province, China. **Current Applied Science And Technology**, p. 453-469, 2020.

FAO. **FAOSTAT:** production sheet. Rome, 2021a.

FAO. **FAOSTAT**: trade sheet. FAO, Rome, 2021b.

FENG, X.; PENG, C.; CHEN, Y.; LIU, X.; FENG, X.; HE, Y. Discrimination of CRISPR/Cas9-induced mutants of rice seeds using near-infrared hyperspectral imaging. **Scientific Reports**, v. 7, n. 1, p. 1-10, 2017. DOI: https://doi.org/10.1038/s41598-017-16254-z.

FOLBERTH, C.; BAKLANOV, A.; BALVIČ, J.; SKALSKÝ, R.; KHABOROV, N; OBERSTEINER, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. **Agricultural and Forest Meteorology**, v. 264, p. 1-15, 2019. DOI: https://doi.org/10.1016/j.agrformet.2018.09.021

FU, J.; FALKE, K. C.; THEIEMANN, A.; SCHRAG, T. A.; MELCHINGER, A. E.; SCHOLTEN, S.; FRISH, M. Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. **Theoretical and Applied Genetics**, v. 124, n. 5, p. 825-833, 2012. DOI: https://doi.org/10.1007/s00122-011-1747-9.

GALLI, G.; ALVEZ, F. C.; MOROSINI, J. S.; FRITSCHE-NETO, R. On the usefulness of parental lines GWAS for predicting low heritability traits in tropical maize hybrids. **PloS one,** v. 15, n. 2, p. e0228724, 2020. DOI: https://doi.org/10.1371/journal.pone.0228724

GOKULNATH, B. V.; GANDHI, U. D. Regularized deep clustering approach for effective categorization of maize diseases. **Journal of Ambient Intelligence and Humanized Computing**, p. 1-10, 2021. DOI: https://doi.org/10.1007/S12652-021-02912-8

GONZÁLEZ-CAMACHO, J. M.; CROSSA, J.; PÉREZ-RODRIGUEZ, P.; ORNELLA, L.; DIANOLA et al. Genome-enabled prediction using probabilistic neural network classifiers. **BMC Genomics**, v. 17, n. 1, p. 1-16, 2016. DOI: https://doi.org/10.1186/s12864-016-2553-1.

GUO, Z.; MAGWIRE, M. M.; BASTEN, C. J.; XU, Z.; WANG, D. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. **Theoretical and Applied Genetics**, v. 129, n. 12, p. 2413-2427, 2016. DOI: https://doi.org/10.1007/s00122-016-2780-5.

GUPTA, S. KUMAR, A., PATEL, R.; KUMAR, V. Genetically modified crop regulations: scope and opportunity using the CRISPR-Cas9 genome editing approach. **Molecular Biology Reports**, p. 1-13, 2021. DOI: https://doi.org/10.1007/s11033-021-06477-9.

HAIGH, T. Actually, Turing did not invent the computer. **Communications of the ACM**, v. 57, n. 1, p. 36-41, 2014. DOI: https://doi.org/10.1145/2542504

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques**. 3. ed. Amsterdam: Elsevier, 2011.

HE, K.; ZHANG, X.; REN, S.; SUM, J. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas. **Proceedings**... New York: IEEE, 2016. p. 770-778.

HE, X.; ZHAO, K.; CHU, X. AutoML: A Survey of the State-of-the-Art. **Knowledge-Based Systems**, v. 212, p. 106622, 2021. DOI: https://dx.doi.org/10.1016/j.knosys.2020.106622

HESLOT, N.; YANG, H. P.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection in plant breeding: a comparison of models. **Crop Science**, v. 52, n. 1, p. 146-160, 2012. DOI: https://doi.org/10.2135/cropsci2011.06.0297.

HOFFMAN, G. E. Correcting for population structure and kinship using the linear mixed model: theory and extensions. **PloS One**, v. 8, n. 10, p. e75707, 2013. DOI: https://doi.org/10.1371/journal.pone.0075707.

HUANG, S.; FAN, X.; SUN, L.; SHEN, Y.; SOU, X. Research on classification method of maize seed defect based on machine vision. **Journal of Sensors**, v. 2019, article 2716975, 2019. DOI: https://doi.org/10.1155/2019/2716975.

GRAHAM, G. J.; GRIFFIN, T. S.; FLEISHER, D. H.; NAUMOVA, E. N.; KOCK, M.; WARDLOW, B. D. Mapping sub-field maize yields in Nebraska, USA by combining remote sensing imagery, crop simulation models, and machine learning. **Precision Agriculture,** v. 21, n. 3, p. 678-694, 2019. DOI: https://doi.org/10.1007/s11119-019-09689-z.

JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**, v. 9, n. 2, p. 166-177, 2010. DOI: https://doi.org/10.1093/bfgp/elq001.

JIANG, J.; XING, F.; WANG, C.; ZENG, X.; ZOU, Q. Investigation and development of maize fused network analysis with multi-omics. **Plant Physiology and Biochemistry**, v. 141, p. 380-387, 2019. DOI: https://doi.org/10.1016/j.plaphy.2019.06.016.

JIANG, L.; BALL, G.; HODGMAN, C.; COULES, A.; ZHAO, H.; LU, C. Analysis of gene regulatory networks of maize in response to nitrogen. **Genes**, v. 9, n. 3, article 151, 2018. DOI: https://doi.org/10.3390/genes9030151.

JIANG, S.; CHENG, Q.; YAN, J.; FU, R.; WANG, X.; et al. Genome optimization for improvement of maize breeding. **Theoretical and Applied Genetics**, v. 133, n. 5, p. 1491-1502, 2020. DOI: https://doi.org/10.1007/s00122-019-03493-z

JIN, S.; SU, Y.; SONG, S.; XU, K.; HU, T.; YANG, Q.; WU, F.; XU, G.; MA, M.; GUAN, H.; MA, Q.; GUAN, H.; PANG, S.; LI, Y.; GUO, Q. Non-destructive estimation of field maize biomass using terrestrial lidar: an evaluation from plot level to individual leaf level. **Plant Methods**, v. 16, p. 1-19, 2020. DOI: https://doi.org/10.1186/s13007-020-00613-5.

KAUL, M.; HILL, R. L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. **Agricultural Systems**, v. 85, n. 1, p. 1-18, 2005. DOI: https://doi.org/10.1016/j.agsy.2004.07.009.

KHAKI, S.; KHALILZADEH, Z.; WANG, L. Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. **Plos One**, v. 15, n. 5, p. e0233382, 2020. DOI: https://doi.org/10.1371/journal.pone.0233382

KHAKI, S.; PHAM, H.; HAN, Y., KUHL; A., KENT, W.; WANG, L. Deepcorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation. **Knowledge-Based Systems,** v. 218, p. 106874, 2021. DOI: https://doi.org/10.1016/j.knosys.2021.106874.

KIM, K. D.; KANG, Y.; KIM, C. Application of genomic big data in plant breeding: Past, present, and future. **Plants**, v. 9, n. 11, p. 1454, 2020. DOI: https://doi.org/10.3390/plants9111454.

KUMAR, V.; MINZ, S. Feature selection: a literature review. **SmartCR**, v. 4, n. 3, p. 211-229, 2014. DOI: https://doi.org/10.6029/smartcr.2014.03.007

KREMLING, Karl A. G. et al. Transcriptome-wide association supplements genome-wide association in *Zea mays*. **G3: Genes, Genomes, Genetics**, v. 9, n. 9, p. 3023-3033, 2019. DOI: https://doi.org/10.1534/g3.119.400549.

LADO, B.; BARRIOS, P. G.; QUINCKE, M.; SILVA, P.; GUTIÉRREZ, L. Modeling genotype x environment interaction for genomic selection with unbalanced data from a wheat breeding program. **Crop Science**, v. 56, n. 5, p. 2165-2179, 2016. DOI: https://doi.org/10.2135/cropsci2015.04.0207.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436-444, 2015. DOI: https://doi.org/10.1038/nature14539

LENG, G.; HALL, J. W. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. **Environmental Research Letters**, v. 15, n. 4, p. 044027, 2020. DOI: https://doi.org/10.1088/1748-9326/ab7b24.

LI, D.; XU, Z.; GU, R.; WANG, P.; LYLE, D.; XU, J.; ZHANG, H.; WANG, G. Enhancing genomic selection by fitting large-effect SNPs as fixed effects and a genotype-by-environment effect using a maize BC1F3: 4 population. **PloS One**, v. 14, n. 10, e0223898, 2019. DOI: https://doi.org/10.1371/journal.pone.0223898.

LI, G., DONG, Y.; ZHAO, Y.; TIAN, X.; WÜRSCHUM, T.; XUE, J.; CHEN, C.; REIF, J. C.; XU, S.; LIU, W. Genome-wide prediction in a hybrid maize population adapted to Northwest China. **The Crop Journal**, v. 8, n. 5, p. 830-842, 2020. DOI: https://doi.org/10.1016/j.cj.2020.04.006.

LIN, J.; WONG, K. C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. **Bioinformatics**, v. 34, n. 17, p. i656-i663, 2018. DOI: https://dx.doi.org/10.1093%2Fbioinformatics%2Fbty554.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A review of machine learning interpretability methods. **Entropy**, v. 23, n. 1, p. 18, 2021. DOI: https://doi.org/10.3390/e23010018.

LIU, H.; SUN, H.; LI, M.; LIDA, M. Application of color featuring and deep learning in maize plant detection. **Remote Sensing**, v. 12, n. 14, p. 2229, 2020a. DOI: https://doi.org/10.3390/rs12142229.

LIU, Y.; CEN, C.; CHE, Y.; KE, R.; MA, Y.; MA, Y. Detection of maize tassels from UAV RGB imagery with faster R-CNN. **Remote Sensing**, v. 12, n. 2, p. 338, 2020b. DOI: https://doi.org/10.3390/rs12020338.

LÓPEZ-CORTÉS, X. A., MATAMALA, F.; MALDONADO, C.; MORA-POBLETE, F.; SCAPIM, C. A. A deep learning approach to population structure inference in inbred lines of maize. **Frontiers in Genetics**, v. 11, p. 1403, 2020. DOI: https://doi.org/10.3389/fgene.2020.543459.

MACKELPRANG, R.; LEMAUX, P. G. Genetic engineering and editing of plants: an analysis of new and persisting questions. **Annual review of plant biology**, v. 71, p. 659-687, 2020. DOI: https://doi.org/10.1146/annurev-arplant-081519-035916

MEJÍA-GUERRA, M. K.; BUCKLER, E. S. A k-mer grammar analysis to uncover maize regulatory architecture. **BMC Plant Biology**, v. 19, n. 1, p. 1-17, 2019. DOI: https://doi.org/10.1186/s12870-019-1693-2.

MELCHINGER, A. E.; GUMBER, R. K. Overview of heterosis and heterotic groups in agronomic crops. **Concepts and Breeding of Heterosis in Crop Plants**, v. 25, p. 29-44, 1998. DOI: https://doi.org/10.2135/cssaspecpub25.c3

MENG, Q. CUI, Z.; YANG, H.; ZHANG, F. CHEN, X. Establishing high-yielding maize system for sustainable intensification in China. **Advances in Agronomy**, v. 148, p. 85-109, 2018. DOI: https://doi.org/10.1016/bs.agron.2017.11.004.

MEUWISSEN, Theo HE; HAYES, Ben J. GODDARD, Michael E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001. DOI: https://doi.org/10.1093/genetics/157.4.1819.

MÔRO, G. V.; SANTOS, M. F.; DE SOUZA JUNIOR, C. L. Comparison of genome-wide and phenotypic selection indices in maize. **Euphytica**, v. 215, n. 4, p. 1-14, 2019. DOI: http://dx.doi.org/10.1007/s10681-019-2401-x.

MUTHAMILARASAN, M.; SINGH, N. K.; PRASAD, M. Multi-omics approaches for strategic improvement of stress tolerance in underutilized crop species: a climate change perspective. **Advances in Genetics**, v. 103, p. 1-38, 2019. DOI: https://doi.org/10.1016/bs.adgen.2019.01.001.

O'BRIEN, A. R.; WILSON, L. O.; BURGIO, G.; BAUER, D. C. Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. **Scientific Reports**, v. 9, n. 1, p. 1-10, 2019. DOI: https://doi.org/10.1038/s41598-019-39142-0.

O'BRIEN, A. R.; BURGIO, G.; BAUER, D. C. Domain-specific introduction to machine learning terminology, pitfalls and opportunities in CRISPR-based gene editing. **Briefings in Bioinformatics**, v. 22, n. 1, p. 308-314, 2021. DOI: https://doi.org/10.1093/bib/bbz145

OLSON, R. S.; LA CAVA, W.; ORZECHOWSKI, P.; URBANOWICZ, R. J.; MOORE, J. H. PMLB: a large benchmark suite for machine learning evaluation and comparison. **BioData Mining**, v. 10, n. 1, p. 1-13, 2017. DOI: https://doi.org/10.1186/s13040-017-0154-4

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. **Computers and Electronics in Agriculture**, v. 74, n. 2, p. 250-257, 2010. DOI: https://doi.org/10.1016/j.compag.2010.08.013.

OSAMA, K.; MISHRA, B. N.; SOMVANSHI, P. Machine learning techniques in plant biology. In: **PlantOmics:** The Omics of Plant Science. Springer, New Delhi, 2015. p. 731-754.

PAZHAMALA, L. T.; KUDAPA, H.; WECKWERTH, W.; MILLAR, A. H.; VARSHNEY, R. K. Systems biology for crop improvement. **The Plant Genome**, v. 14, n. 2, e20098, 2021. DOI: https://doi.org/10.1002/tpg2.20098.

PÉREZ-RODRÍGUEZ, P.; FLORES-GALARZA, S.; VAQUERA-HUERTA, H.; DEL VALLE-PANIAGUA, D. H.; MONTESINOS-LÓPEZ, O. A.; CROSSA, J. Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. **The Plant Genome**, v. 13:e20021, p. 1-13, 2020. DOI: https://doi.org/10.1002/tpg2.20021

POOK, T.; FREUDENTHAL, J.; KORTE, A.; SIMIANER, H. Using local convolutional neural networks for genomic prediction. **Frontiers in Genetics**, v. 11, p. 1366, 2020. DOI: https://doi.org/10.3389/fgene.2020.561497

PRASAD, S. S.; SUGANESH, R.; THANGATAMILAN, M. Machine Learning Approach for Crop Prediction Based on Climatic Parameters. In: **Materials, Design, and Manufacturing for Sustainable Environment**. Springer, Singapore, 2021. p. 729-738.

PRESSOIR, G.; BERTHAUD, J. Patterns of population structure in maize landraces from the Central Valleys of Oaxaca in Mexico. **Heredity**, v. 92, n. 2, p. 88-94, 2004a. DOI: https://doi.org/10.1038/sj.hdy.6800387.

PRESSOIR, G.; BERTHAUD, J. Population structure and strong divergent selection shape phenotypic diversification in maize landraces. **Heredity**, v. 92, n. 2, p. 95-101, 2004b. DOI: https://doi.org/10.1038/sj.hdy.6800388.

PROHENS, J. Plant breeding: a success story to be continued thanks to the advances in genomics. **Frontiers in Plant Science**, v. 2, p. 51, 2011. DOI: https://doi.org/10.3389/fpls.2011.00051.

QI, W.; ZHU, T.; TIAN, Z.; LI, C.; ZHANG, W.; SONG, R. High-efficiency CRISPR/Cas9 multiplex gene editing using the glycine tRNA-processing system-based strategy in maize. **BMC Biotechnology**, v. 16, n. 1, p. 1-8, 2016. DOI: https://doi.org/10.1186/s12896-016-0289-2.

QIU, Z.; CHENG, Q.; SONG, J.; TANG, Y.; MA, C. Application of machine learning-based classification to genomic selection and performance improvement. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING, 2016, Lanzhou. **Proceedings**... Cham: Springer, 2016. p. 412-421.

RACHMATIA, H.; KUSUMA, W. A.; HASIBUAN, L. S. Prediction of maize phenotype based on whole-genome single nucleotide polymorphisms using deep belief networks. In: **Journal of Physics: Conference Series**. IOP Publishing, 2017. p. 012003. DOI: https://doi.org/10.1088/1742-6596/835/1/012003.

RAI, A. Explainable AI: from black box to glass box. **Journal of the Academy of Marketing Science**, v. 48, n. 1, p. 137-141, 2020. DOI: https://doi.org/10.1007/s11747-019-00710-5

RAMSTEIN, G. P.; JENSEN, S. E.; BUCKLER, E. S. Breaking the curse of dimensionality to identify causal variants in Breeding 4. **Theoretical and Applied Genetics**, v. 132, n. 3, p. 559-567, 2018. DOI: https://doi.org/10.1007/s00122-018-3267-3.

RANUM, P.; PEÑA-ROSAS, J. P.; GARCIA-CASAL, M. N. Global maize production, utilization, and consumption. **Annals of the New York Academy of Sciences**, v. 1312, n. 1, p. 105-112, 2014. DOI: https://doi.org/10.1111/nyas.12396

RAWAT, W.; WANG, Z. Deep convolutional neural networks for image classification: A comprehensive review. **Neural Computation**, v. 29, n. 9, p. 2352-2449, 2017. DOI: https://doi.org/10.1162/neco_a_00990.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. **Advances in Neural Information Processing Systems**, v. 28, p. 91-99, 2015.

RICE, B.; LIPKA, Al. E. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. **The Plant Genome**, v. 12, n. 1, 2019. DOI: https://doi.org/10.3835/plantgenome2018.07.0052.

RIEDELSHEIMER, C.; TECHNOW, F.; MELCHINGER, A. E; Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. **BMC Genomics**, v. 13, n. 1, p. 1-9, 2012. DOI: https://doi.org/10.1186/1471-2164-13-452.

RIPPEY, Bradley R. The US drought of 2012. **Weather and Climate Extremes**, v. 10, p. 57-64, 2015. DOI: https://doi.org/10.1016/j.wace.2015.10.004

SARIJALOO, F. B.; PORTA, M.; TASLIMI, B.; PARDALOS, P. M. Yield performance estimation of corn hybrids using machine learning algorithms. **Artificial Intelligence in Agriculture**, v. 5, p. 82-89, 2021. DOI: http://dx.doi.org/10.1016/j.aiia.2021.05.001

SCHAEFER, R. J.; MICHNO, J. M.; JEFFERS, J.; HOEKENGA, O.; DILKES, B.; BAXTER, I.; MYERS, C. L. Integrating coexpression networks with GWAS to prioritize causal genes in maize. **The Plant Cell**, v. 30, n. 12, p. 2922-2942, 2018. DOI: https://doi.org/10.1105/tpc.18.00299.

SCHAEFER, R. J.; MICHNO, J. M.; JEFFERS, J.; HOEKENGA, O.; DILKES, B.; BAXTER, I.; MYERS, C. L. Maize yield and nitrate loss prediction with machine learning algorithms. **Environmental Research Letters**, v. 14, n. 12, p. 124026, 2019. DOI: https://doi.org/10.1088/1748-9326/ab5268.

SHAKOOR, N.; LEE, S.; MOCKLER, T. C. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. **Current Opinion in Plant Biology**, v. 38, p. 184-192, 2017. DOI: https://doi.org/10.1016/j.pbi.2017.05.006

SHETE, S.; SRINIVASAN, S.; GONSALVES, T. A. TasselGAN: An Application of the Generative Adversarial Model for Creating Field-Based Maize Tassel Data. **Plant Phenomics**, v. 2020, 2020. DOI: https://doi.org/10.34133/2020/8309605

SIBIYA, M.; SUMBWANYAMBE, M. Automatic Fuzzy Logic-Based Maize Common Rust Disease Severity Predictions with Thresholding and Deep Learning. **Pathogens**, v. 10, n. 2, p. 131, 2021. DOI: https://doi.org/10.3390/pathogens10020131

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2015, San Diego. **Proceedings**… San Diego: [s.n.], 2015. Available in: https://arxiv.org/pdf/1409.1556.pdf. Access in: 13 nov. 2021.

SINGH, R. P.; CHINTAGUNTA, A. D.; AGARWAL, D. K.; KUREEL, R. S.; KUMAR, S. J. Varietal replacement rate: Prospects and challenges for global food security. **Global Food Security**, v. 25, p. 100324, 2020. DOI: https://doi.org/10.1016/j.gfs.2019.100324.

SUN, J.; YANG, Y.; HE, X.; WU, X. Northern maize leaf blight detection under complex field environment based on deep learning. **IEEE Access**, v. 8, p. 33679-33688, 2020. DOI https://doi.org/10.1109/ACCESS.2020.2973658.

SULTANA,F.;SUFIAN,A.;DUTTA,P.Advancements in image classification using convolutional neural network. In: INTERNATIONAL CONFERENCE ON RESEARCH IN COMPUTATIONAL INTELLIGENCE AND COMMUNICATION NETWORKS, 4., 2018. **Proceedings**… New York: IEEE, 2018. p. 122-129.

SVITASHEV, S.; YOUNG, J. K.; SCHWARTZ, C.; GAO, H.; FALCO, S. C.; CIGAN, A. M. Targeted mutagenesis, precise gene editing, and site-specific gene insertion in maize using Cas9 and guide RNA. **Plant Physiology**, v. 169, n. 2, p. 931-945, 2015. DOI: https://doi.org/10.1104/pp.15.00793.

SWARUP, S.; CARGILL, E. J.; CROSBY, K.; FLAGEL, L.; KNISKERN, J.; GLENN, K. C. Genetic diversity is indispensable for plant breeding to improve crops. **Crop Science**, v. 61, n. 2, p. 839-852, 2021. DOI: https://doi.org/10.1002/csc2.20377

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2015, Boston. **Proceedings**… New York: IEEE, 2015. p. 1-9.

TARCA, A. L.; CAREY, V. J.; CHEN, X.; ROMERO, R.; DRĂGHICI, S. Machine learning and its applications to biology. **PLoS Computational Biology**, v. 3, n. 6, p. e116, 2007. DOI: https://doi.org/10.1371/journal.pcbi.0030116.

TIAN, Z.; WANG, J. W.; LI, J.; HAN, B. Designing future crops: challenges and strategies for sustainable agriculture. **The Plant Journal**, v. 105, n. 5, p. 1165-1178, 2021. DOI: https://doi.org/10.1111/tpj.15107.

TONG, H.; NIKOLOSKI, Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. **Journal of Plant Physiology**, v. 257, p. 153354, 2021. DOI: https://doi.org/10.1016/j.jplph.2020.153354

TURING, A. M. On computable numbers, with an application to the Entscheidungs problem. **Proceedings of the London Mathematical Society**, v. 2, n. 1, p. 230-265, 1937.

VAN DIJK, A. D. J.; KOOTSTRA, G.; KRUIJER, W.; DE RIDDER, D. Machine learning in plant science and plant breeding. **Iscience**, p. 101890, 2020. DOI: https://doi.org/10.1016/j.isci.2020.101890.

VARGAS, R.; MOSAVI, A.; RUIZ, R. **Deep learning**: a review. Advances in Intelligent Systems and Computing, 2017. DOI: https://doi.org/10.20944/PREPRINTS201810.0218.V1.

VERGARA-DÍAZ, O.; ZAMAN-ALLAH, M. A.; MASUKA, B.; HORNERO, A.; ZARCO-TEJADA, P.; PRASANNA, B. M.; CAIRNS, J. E; ARAUS, J. L. A novel remote sensing approach for prediction of maize yield under different conditions of nitrogen fertilization. **Frontiers in Plant Science**, v. 7, p. 666, 2016. DOI: https://dx.doi.org/10.3389%2Ffpls.2016.00666.

VIGOUROUX, Y.; GLAUBITZ, J. C.; MATSUOKA, Y.; GOODMAN, M. M.; SÁNCHEZ G, J.; DOEBLEY, J. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. **American Journal of Botany**, v. 95, n. 10, p. 1240-1253, 2008. DOI: https://doi.org/10.3732/ajb.0800097

VINODKUMAR, P. K.; OZCINAR, C.; ANBARJAFARI, G. Prediction of sgRNA Off-Target Activity in CRISPR/Cas9 Gene Editing Using Graph Convolution Network. **Entropy**, v. 23, n. 5, p. 608, 2021. DOI: https://doi.org/10.3390/e23050608

VOSS-FELS, K.; COOPER, M.; HAYES, B. J. Accelerating crop genetic gains with genomic selection. **Theoretical and Applied Genetics**, v. 132, n. 3, p. 669-686, 2019. DOI: https://doi.org/10.1007/s00122-018-3270-8.

WALLACE, J. G.; RODGERS-MELNICK, E.; BUCKLER, E. S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. **Annual Review of Genetics**, v. 52, p. 421-444, 2018. DOI: https://doi.org/10.1146/annurev-genet-120116-024846.

WANG, J.; XIANG, X.; BOLUND, L.; ZHANG, X.; CHENG, L.; LUO, Y. GNL-Scorer: a generalized model for predicting CRISPR on-target activity by machine learning and featurization. **Journal of Molecular Cell Biology**, v. 12, n. 11, p. 909-911, 2020. DOI: https://doi.org/10.1093/jmcb/mjz116.

WASHBURN, J. D.; BURCH, M. B.; FRANCO, J.; VALDES, A.. Predictive breeding for maize: Making use of molecular phenotypes, machine learning, and physiological crop models. **Crop Science**, v. 60, n. 2, p. 622-638, 2020. DOI: https://doi.org/10.1002/csc2.20052.

WENG, Z. From conventional machine learning to AutoML. In: **Journal of Physics: Conference Series**. IOP Publishing, 2019. p. 012015. DOI: http://dx.doi.org/10.1088/1742-6596/1207/1/012015.

WHEELER, T.; VON BRAUN, J. Climate change impacts on global food security. **Science**, v. 341, n. 6145, p. 508-513, 2013. DOI: https://doi.org/10.1126/science.1239402

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

XIONG, H., LIU, D., LI, Q., LEI, M., XU, L., WU, L., WNAG, Z.; REN, S.; LI, W.; XIA, M.; LU, L.; LU., H., HOU, Y.; ZHU, S.; LIU, XIN.; SUN, Y.; WANG, J.; YANG, H.; WU, K.; XU, X.; LEE, L. J. RED-ML: a novel, effective RNA editing detection method based on machine learning. **Gigascience**, v. 6, n. 5, p. gix012, 2017. DOI: https://doi.org/10.1093/gigascience/gix012.

XUE, L.; TANG, B.; CHEN, W.; LUO, J. Prediction of CRISPR sgRNA activity using a deep convolutional neural network. **Journal of Chemical Information and Modeling**, v. 59, n. 1, p. 615-624, 2018. DOI: https://doi.org/10.1021/acs.jcim.8b00368

YANG, W.; FENG, H.; ZHANG, X.; ZHANG, J.; DOONAN, J. H.; BATCHELOR, W. D.; XIONG, J.; YAN, J. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. **Molecular Plant**, v. 13, n. 2, p. 187-214, 2020. DOI: https://doi.org/10.1016/j.molp.2020.01.008.

YIN, L.; ZHANG, H.; ZHOU, X.; YUAN, X.; ZHAO, S.; LI, X.; LIU, X. KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. **Genome Biology**, v. 21, n. 1, p. 1-22, 2020. DOI: https://doi.org/10.1186/s13059-020-02052-w.

YU, H.; LI, J. Short-and long-term challenges in crop breeding. **National Science Review**, v. 8, n. 2, p. nwab002, 2021. DOI: https://doi.org/10.1093/nsr/nwab002

ZHANG, C.; ZHAO, Y.; YAN, T.; BAI, X.; XIAO, Q.; GAO, P.; LI. M.; HUANG, W.; BAO, Y.; LIU, F. Application of near-infrared hyperspectral imaging for variety identification of coated maize kernels with deep learning. **Infrared Physics & Technology**, v. 111, p. 103550, 2020a. DOI: https://doi.org/10.1016/j.infrared.2020.103550

ZHANG, L.; ZHANG, Z.; LUO, Y., CAO, J.; TAO, F. Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches. **Remote Sensing**, v. 12, n. 1, p. 21, 2020b. DOI: https://doi.org/10.3390/rs12010021.

ZHAO, Y.; GOWDA, M.; LIU, W.; WÜRSCHUM, T.; MAURER, H.; P.; LONGIN, F, H.; RANC, N.; REIF, J. C. Accuracy of genomic selection in European maize elite breeding populations. **Theoretical and Applied Genetics**, v. 124, n. 4, p. 769-776, 2012. DOI: https://doi.org/10.1007/s00122-011-1745-y.

ZHOU, P.; LI, Z.; MAGNUSSON, E.; GOMEZ CANO, F.; CRISP, P. A.; NOSHAY, J. M.; GROTEWOLD E.; HIRSCH, C.; BRIGGS.; S. P.; SPRINGER, N. M. Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. **The Plant Cell**, v. 32, n. 5, p. 1377-1396, 2020. DOI: https://doi.org/10.1105/tpc.20.00080

ZHOU, S.; CHAI, X.; YANG, Z.; WANG, H.; YANG, C.; SUN, T. Maize-IAS: a maize image analysis software using deep learning for high-throughput plant phenotyping. **Plant Methods**, v. 17, n. 1, p. 1-17, 2021. DOI: https://doi.org/10.1186/s13007-021-00747-0

ZIMMER, L.; LINDAUER, M.; HUTTER, F. Auto-Pytorch: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2021. DOI: https://doi.org/10.1109/TPAMI.2021.3067763

ZINGARETTI, M. L.; MONFORT, A.; PÉREZ-ENCISO, M. pSBVB: a versatile simulation tool to evaluate genomic selection in polyploid species. G3: **Genes, Genomes, Genetics**, v. 9, n. 2, p. 327-334, 2019. DOI: https://doi.org/10.1534/g3.118.200942