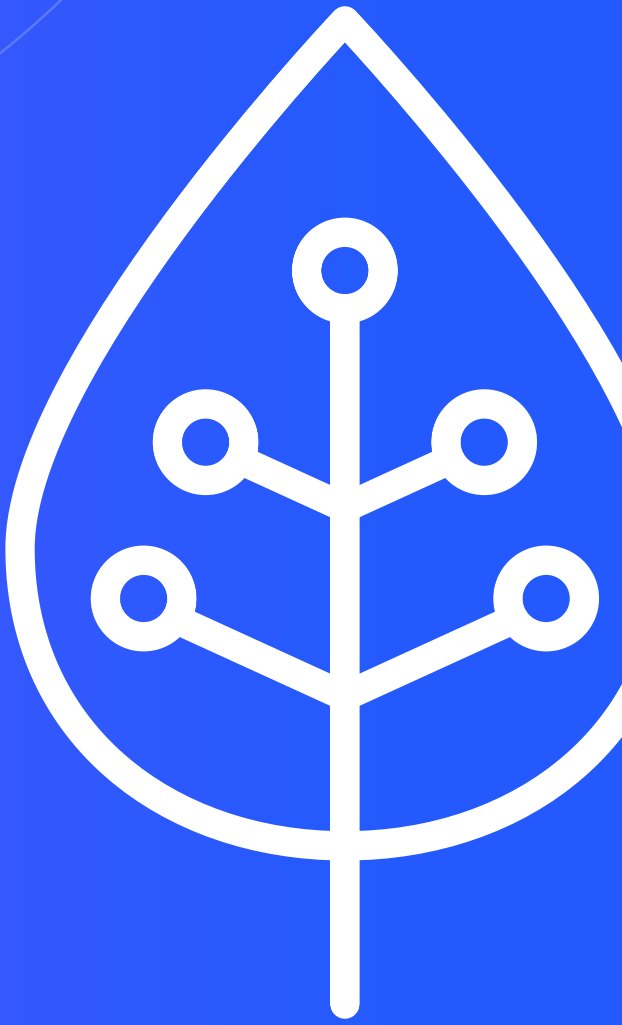
Benchling

How Artificial Intelligence Is Advancing Plant Biology



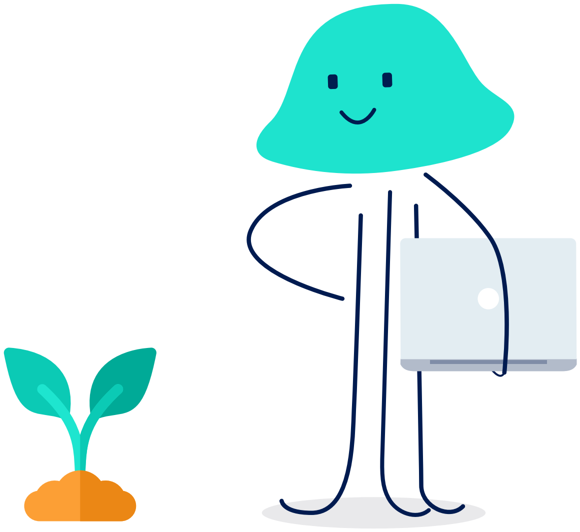
Benchling's overview
of recent research at
the cutting edge of AgBio

Table of Contents

How Artificial Intelligence Is Advancing Plant Biology

To meet the pressing challenges of a changing climate and growing population, plant biologists, breeders, and farmers are harnessing a slew of new technologies. From the lab, to the greenhouse, to the field, advances in next generation sequencing (NGS), wireless sensors, and unmanned aerial vehicles are accelerating discovery and reducing the time from idea to commercial implementation. The critical capability that enables scientists to make sense of the wealth of data generated by these technologies and design effective strategies is artificial intelligence (AI).

As a partner to leading agricultural biotechnology companies, Benchling is proud to help customers manage lab operations and information, so that their data is ready to fuel critical AI-driven insights. In this white paper, we celebrate recent advances that AI has brought to agricultural plant biology.



Automating labor intensive processes

In the past year we've seen a number of papers where AI is helping plant biologists automate labor intensive processes, especially in planted fields. In these applications, AI is used primarily to automate image analysis for biovolume measurement [1, 2], identification of plant parts such as maize tassels [3], biochemical parameter measurement [4, 5, 6] or disease detection [7-10] in the field, and to screen seeds and seedlings for desirable qualities [6, 11], germination [12], or disease [13].

Goals

While each of these studies has different ultimate end-goals, the use of AI is meant to help growers make better-informed decisions as quickly as possible, whether it's to compare the growth of different strains and different conditions in the field, to find the most promising seeds or seedlings to plant, or to identify diseased plants as early as possible so that outbreaks can be prevented with fewer plants needing to be destroyed.

Challenges and solutions

These studies highlight two different types of challenges for AI and machine vision, namely image segmentation and object classification.

Image segmentation: Through a combination of more affordable camera and data storage capabilities with well-developed machine learning (ML) algorithms, scientists are expanding automated image analysis from the well-controlled conditions of the lab to the highly variable conditions in the field. Because the goal is to ultimately create a system that can be used by growers using their own cameras, potentially even phone cameras, the analysis needs to be able to handle images taken at multiple angles, multiple heights, and under different environmental conditions.

To meet these challenges, the majority of studies expanded on the "regions with convolutional neural networks" (R-CNN) approach to increase the accuracy of pixel assignment and object identification [1, 3, 9, 10]. Colorado, *et al* [2], took a semi-manual image segmentation algorithm, GrabCut, and removed the manual component by using an unsupervised ML approach—a Montecarlo-sampled K-means segmentation—to create two binary masks to segment the image with further refinement using Guided Filtering.

Object classification: A number of the image analysis studies bypassed the problem of image segmentation by using either manual or semi-manual segmentation methods [6, 7, 8]. Others relied on spectral data [4, 5], which does not require image segmentation.

Neural network approaches were used in many of these papers, with Sabanci [13] using artificial bee colony (ABC) optimization in artificial neural networks (ANN) and extreme learning machine algorithms to detect sunn pest disease in wheat kernels, Boulent, *et al* [7] using convolutional neural networks and fully convolutional neural networks to detect flavescence dorée symptoms in grapevines, and Sibiya and Sumbwanyambe [8] moving beyond disease detection to classify plants by disease severity using a convolutional neural networks algorithm.

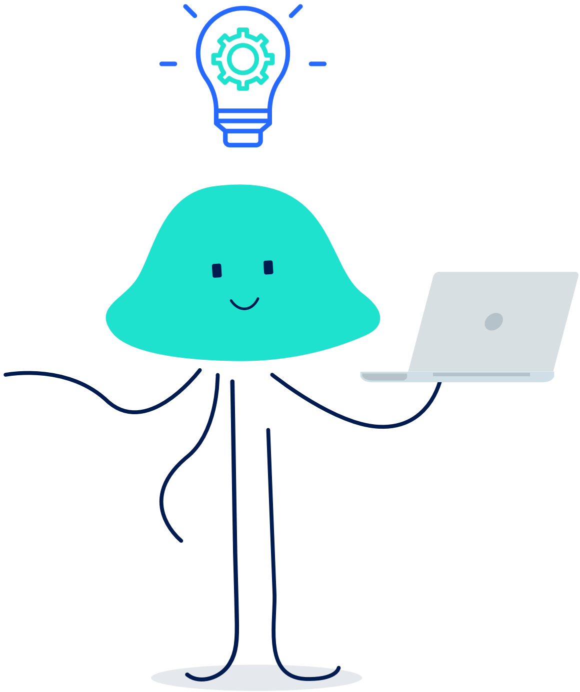
Ahmed, *et al* [6], combined X-ray analysis of seedlings with different deep learning approaches to develop a model that could predict successful germination based on the internal morphology of the seed. They compared deep learning approaches—simple ConvNet, AlexNet, VGG-19, ResNet-50, and ResNet-101—to more conventional classifiers and found that the ResNet-50 outperformed all other classifiers with a predictive accuracy of 87.3%.

Siedliska, *et al* [4], asked if ML analysis of hyperspectral imaging data can differentiate plants from four different phosphorus (P) treatments. They compared backpropagation neural networks, random forests, naive Bayes, and support vector machines (SVM) approaches and obtained the best predictions using a random forests approach.

D'Odorico, *et al* [5], developed an ML model using a random forests approach that was able to rapidly assess plant stress (Fv/Fm) in 6000 conifer seedlings.

Both image segmentation and object classification: Masteling, *et al* [12], developed an automated assay to screen for potential herbicides by discriminating between germinated and ungerminated seeds of the parasitic weed Striga. This study used YOLOv3 for both image segmentation and classification and chose this method over others due to its ability to be used on computers without a specialized GPU.

Another study [11] demonstrated the accessibility of ML by using the free interactive ML toolkit Ilastik [14], which performs image segmentation based on a random forests classifier, to identify soybean seeds and seedlings in images, classify seeds based on appearance, identify vigorously growing seedlings, and correlate seed appearance with growth potential.



Powering genomic selection/genomic prediction

Genomic selection (GS) holds the promise of accelerating crop improvement by helping breeders more accurately identify the crosses that will yield the desired traits based on data from genome-wide genotyping. To achieve the high marker density needed for GS, genotyping is typically done using next generation sequencing (NGS), an approach called genotyping by sequencing (GBS).

GS/genomic prediction (GP) typically relies on linear or Bayesian methods and it is unclear how much benefit AI can provide over conventional methods. Recent published studies do not resolve this question, although they add more examples of where AI can enhance prediction accuracy, as well as examples where it does not.

Goals

Ultimately, the use of GS should lead to faster, cheaper, and less labor-intensive crop improvement than conventional breeding methods. Whether AI can improve the accuracy of GP compared to conventional methods remains to be seen, although at this point the question is more likely to be “under which conditions can AI improve GP accuracy,” rather than simply “should AI be used for GP?”

Challenges and solutions

While GS has positively impacted livestock breeding, the cost versus benefit equation has lagged for agricultural crops [15]. One reason is the higher intrinsic value of individuals in livestock breeding. Two more are the added complexities that spring from the polyploid genomes possessed by many agriculturally important crops as well as the highly variable interaction between genetics and the environment that impacts crops more than livestock.

To increase the accuracy of GP, many agricultural researchers are exploring different AI-based approaches [16, 17-20, 21-23, 24]. For example, Toda, *et al* [24], used standard linear regression methods for GP and then explicitly accounted for environmental effects by integrating either a deterministic crop growth model or an ML approach using a random forests algorithm. Both deterministic and ML approaches delivered modest increases in biomass prediction accuracy over the conventional linear methods.

That said, the research in 2020 has not resolved the question of how useful or when to use AI for GS. Several papers found that conventional linear methods outperformed ML and deep learning (DL) approaches [20,22-23]. Zingaretti, *et al* [22], found that convolutional neural networks (CNNs) outperformed conventional approaches only in situations with strong epistasis, and Pook, *et al* [20], who used localized CNNs, found that ML could not outperform linear models when the training set was small.

Interestingly, the papers that did show an improvement in prediction accuracy relied on different algorithms. While the specifics of each study—the traits and species assessed—are critical to the analysis, it's worth noting which approaches were successful, namely random forests [18, 21, 24], multilayer perceptron (MLP) [16], Bayes regularized neural networks [19], SVM [21], and an ensemble approach that used support vector regression (SVR), kernel ridge regression, and elastic net [17].

Understanding environmental impacts

One way to increase the power of GS is to build larger phenomic and envirotypic databases, and integrate these databases with genotype information. In the past, building phenomic and envirotypic databases has been labor intensive. However, the lower costs and wider access of machine vision, multispectral imaging, mobile phone cameras, and drone technologies is making daily environmental and phenotypic measurements possible with the help of AI [25].

In addition, AI approaches are being used in studies where extensive measurements already exist and the goal is to understand which parameters are most predictive of growth in response to different environmental conditions [26,27].

Goals

Here, AI is used to enhance understanding when processes are highly complex and governed by multiple parameters, providing insight into which measurements are most predictive of the desired trait or how different parameters interact to affect the desired trait.

Challenges and solutions

As discussed above, GS is challenged by the lack of detailed information on environmental conditions. Jin, *et al* [26], demonstrate just how powerful that information can be by leveraging a highly detailed set of data collected over ten years on almond orchards in California. Taking into account light interception, climate information, almond yield, cultivar, and other environmental and crop management conditions, they built a model using a random forests algorithm that was able to explain 82% of yield variation, with ~60% of that variability due to changes in light interception.

Coulibali, *et al* [27], use a similar approach with potato, leveraging data dating back to 1979 to build a predictive model that could explain how fertilization regime—nitrogen, phosphorus, potassium—interact with weather, soils, and land management impact tuber yield and quality (defined as size and specific gravity). Overall, their models had moderate predictive ability (R2 > 0.7), with Gaussian processes algorithms providing the best performance.

Yoosefzadeh-Najafabadi, *et al* [25], tackled a different problem, predicting soybean yield based on hyperspectral measurements of plants at an early stage of development, to help breeders choose promising plants earlier in the strain development process. They used an ensemble stacking approach that incorporated random forest, MLP, and SVM, and achieved 87 - 93% prediction accuracy, depending on the number of variables used.

And Saffariha, *et al* [28], generated data on Salvia limbata seed germination under different temperature, draught, salinity, and pH conditions to build a predictive model. They found that the MLP algorithm performed better than a linear regression model, with R2 > 0.9.

Optimizing *in vitro* plant culture and transformation

Throughout history, plants have been used medicinally in addition to providing basic nutritional requirements. An increasing number of agricultural biotechnology companies are turning to in vitro plant culture to shorten improvement times or even to bypass field growth. However, developing a robust in vitro plant culture process can be challenging due to the complex interactions of hormones and minerals. A number of researchers are using AI to help identify optimum growth conditions [29, 30, 31], to maximize secondary metabolite production [32], and to optimize Agrobacterium-mediated genetic transformation [33].

Goals

AI is being used to reduce the time it takes to develop optimal conditions for in vitro plant culture, secondary metabolite production, and genetic transformation with agrobacteria.

Challenges and solutions

Optimizing growth: Both Garcia-Perez, *et al* [29], and Hameg, *et al* [30], use ANN and neurofuzzy logic (FormRules® v.4.03 software (Intelligensys LTD, UK)) to generate easy to understand "IF-THEN" statements that relate a specific factor with a response. Hameg, *et al* [30], also build on this information and use ANN combined with genetic algorithms (INForm® v5.01 (Intelligensys Ltd., United Kingdom)) to predict new mineral media formulations that also significantly enhanced growth of their target species, kiwiberry (Actinidia arguta).

Hesami, *et al* [31], compare two different ML methods for optimizing culture conditions for chrysanthemum embryogenesis, MLP and SVR. They found the best predictive accuracy with SVR using non-dominated sorting genetic algorithm-II (NSGA-II) and achieved a somatic embryogenesis rate of 99.09%.

Maximizing secondary metabolite production: Salehi, *et al* [32], explored the use of a general regression neural network-fruit fly optimization algorithm (GRNN-FOA) and a MLP-genetic algorithm (MLP-GA) to find optimal conditions for the production of paclitaxel, an anticancer agent, in Corylus avellana. Fungal elicitors can upregulate production of paclitaxel in in vitro culture, and Salehi, *et al* [32], wanted to quickly find the combination of four input conditions that would maximize paclitaxel production. In their studies, the GRNN-FOA provided the best concordance between predicted and experimental growth data.

Optimizing Agrobacterium-mediated transformation: Just as with the earlier studies searching for optimal growth media and secondary metabolite production, AI is proving to be a valuable tool for optimizing Agrobacterium-mediated genetic transformation. Hesami, *et al* [33], evaluated three ML approaches individually—MLP, adaptive neuro-fuzzy inference system (ANFIS), and radial basis function (RBF)—and as a combined ensemble linked to a fruit fly optimization algorithm (FOA) to optimize Agrobacterium-mediated transformation of chrysanthemum based on 11 input conditions. As one might expect, the ensemble-FOA performed better than each method alone, with a maximum transformation efficiency of 37.54%.

Conclusion

AI is delivering tremendous benefits for agricultural biotech researchers, rapidly assessing the phenotype of individuals in planted fields, accelerating genetic gain, disambiguating the effects of different environmental and genetic parameters on growth, and accelerating optimization of complex, multi-input processes such as *in vitro* growth and bacterial transformation.

Benchling is proud to help teams like these manage their data and streamline the preparation and transfer of critical information from the Benchling Life Sciences R&D Cloud to the AI analysis platform and back. With Benchling's expert implementation teams and cloud-based solution, Benchling can help any agricultural biotech lab unify much of their current lab operations into a single software platform that efficiently and effectively feeds into any AI analysis pipeline.

A Unified Platform for a New Era in Life Sciences

Benchling is a cloud-based platform that supports your entire R&D process from early research to late-stage development by delivering a wide range of functionalities, flexible integration and configuration, and user-friendly, easy-to-navigate software that is compatible with your R&D workflows.

With Benchling, you get a seamless solution for sample tracking, inventory management, regulatory compliance, instrument integration, process management, and reporting and analytics. All of these applications share the same configurable data model, so they can be accessed through a single database—making teams more connected and productive.

Benchling Applications:

Notebook: The most user-friendly electronic lab notebook in the industry — reduce time to data entry by 85% with the first cloud-based notebook built for modern life science.

Molecular Biology: A comprehensive molecular biology suite of 10+ tools in one collaborative environment—build, share, and record DNA and amino acid sequences in one unified platform.

Registry: A convenient and easy-to-use registration system built from the ground up for large molecule R&D—model biological entities and easily enter and extract the data you need.

Inventory: Laboratory inventory management—track the locations of vials, wells, batches, and more, and automatically link results to them.

Requests: A single system for request submission and fulfillment—track request progress in real-time and optimize resource allocation.

Workflows: A digital process management platform for early research through development—organize any scientific workflow into trackable stages and test new variations to optimize your outputs.

Insights: Easy-to-configure dashboards and reports provide a full view of R&D—everyone from scientists and managers to executives can have customized dashboards to deliver the answers that *they* need.

Benchling Platform:

Lab Automation: Synchronize informatics and instrumentation to drive efficiency and maximize the potential of your full suite of molecular laboratory tools.

Validated Cloud: Unify R&D teams for easier collaboration and data management to accelerate time-to-market while meeting compliance requirements with a modern cloud architecture and enterprise-grade security.

Codeless Configuration: Configure and reconfigure workflows without a single line of code so you can quickly adapt and align to your evolving R&D process, resulting in rapid prototyping, full deployment, and fast time to value.

Developer Platform: Integrate any instrument, software, and/or database with Benchling to centralize your data and enable automation. Equipped with APIs, Events, and Data Warehouse, so you can connect Benchling to existing IT infrastructure with ease.

> "We can look at a plant and simultaneously know not only the plasmid we used, but it's exact sequence and dozens of aspects about it. This takes us from using Benchling as a plant inventory system to Benchling as a fully integrated biotechnology pipeline."
>
> Nathaniel Graham, Molecular Biology Scientist, Pairwise Plants

Join over 300,000 scientists using Benchling to power life science R&D.
Request a Demo.

References

Citation		AI Approach	Species
1	Safanova A, *et al.* Olive Tree Biovolume from UAV Multi-Resolution Image Segmentation with Mask R-CNN. Sensors (Basel). 2021 Feb 25;**21**(5):1617. PMCID: PMC7956790.	Mask R-CNN	Olive
2	Colorado JC, *et al.* A novel NIR-image segmentation method for the precise estimation of above-ground biomass in rice crops. *PLoS One.* 2020; **15**(10): e0239591. PMCID: PMC7535130.	GrabCut, Montecarlo-sampled K-means segmentation, guided filtering	Rice
3	Zou H, *et al.* Maize tassels detection: a benchmark of the state of the art. *Plant Methods.* 2020; **16**:108. PMCID: PMC7414585.	Faster R-CNN, YOLOv3, FaceBoxes, RetinaNet, TasselNet	Maize
4	Siedliska A, *et al.* Identification of plant leaf phosphorus content at different growth stages based on hyperspectral reflectance. *BMC Plant Biol.* 2021; **21**: 28. PMCID: PMC7792193.	Backpropagation Neural Network, Random Forests, Naive Bayes and SVM	Celery, Sugar Beet, Strawberry
5	D'Odorico P, *et al.* High-throughput drone-based remote sensing reliably tracks phenology in thousands of conifer seedlings. *New Phytol.* 2020 Jun; **226**(6):1667-1681. PMID: 32157698.	Random forests	White spruce
6	Ahmed MR, *et al.* Classification of Watermelon Seeds Using Morphological Patterns of X-ray Imaging: A Comparison of Conventional Machine Learning and Deep Learning. *Sensors* (Basel). 2020 Dec; **20**(23): 6753. PMCID: PMC7731397.	ConvNet, AlexNet, VGG-19, ResNet-50, and ResNet-101	Watermelon
7	Boulent J, *et al.* utomatic Detection of Flavescence Dorée Symptoms Across White Grapevine Varieties Using Deep Learning. *Front Artif Intell.* 2020; **3**: 564878. PMCID: PMC7944144.	CNN, fully convolutional networks	Grape
8	Sibiya M and Sumbwanyambe M. Automatic Fuzzy Logic-Based Maize Common Rust Disease Severity Predictions with Thresholding and Deep Learning. *Pathogens.* 2021 Feb; **10**(2): 131. PMCID: PMC7912646.	VGG-16 (CNN)	Maize
9	Ahmad J, *et al.* Disease Detection in Plum Using Convolutional Neural Network under True Field Conditions. *Sensors* (Basel). 2020 Oct; **20**(19): 5569. PMCID: PMC7583767.	Inception-v3 (CNN)	Plum
10	Akagi T, *et al.* Explainable Deep Learning Reproduces a 'Professional Eye' on the Diagnosis of Internal Disorders in Persimmon Fruit. *Plant Cell Physiol.* 2020 Dec 23; **61**(11):1967-1973. PMID: 32845307.	VGG16, InceptionsV3, ResNet50 and InceptionResnetV2 (CNNs), Guided backpropagation, Grad-CAM and Guided Grad-CAM	Persimmon
11	de Medeiros AD, *et al.* Interactive machine learning for soybean seed and seedling quality classification. *Sci Rep.* 2020; **10**:11267. PMCID: PMC7347887.	Random forests (Ilastik), linear discriminant analysis, SVM	Soybean
12	Masteling R, *et al.* DiSCount: computer vision for automated quantification of Striga seed germination. *Plant Methods.* 2020 May 1; **16**:60. PMCID: PMC7195706.	YOLO (fully CNN)	Striga
13	Sabanci K. Detection of sunn pest-damaged wheat grains using artificial bee colony optimization-based artificial intelligence techniques Kadir Sabanci. *J Sci Food Agric.* 2020 Jan 30;**100**(2):817-824. PMID: 31646637.	Artificial bee colony (ABC) optimization-based ANN, extreme learning machine (ELM)	Wheat
14	Sommer, C. *et al.* Ilastik: Interactive learning and segmentation toolkit. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro.* Chicago, IL, USA, 2011, pp. 230-233, doi: 10.1109/ISBI.2011.5872394.	NA	NA
15	Xu Y, *et al.* Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Commun.* 2019 Oct 16; **1**(1):100005. PMCID: PMC7747995.	NA	NA

Citation		AI Approach	Species
16	Sandhu KS, *et al*. Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. *Front Plant Sci*. 2020; **11**:613325. PMCID: PMC7813801.	MLP, CNN	Wheat
17	Liang M, *et al*. A Stacking Ensemble Learning Framework for Genomic Prediction. *Front Genet*. 2021; **12**:600040. PMCID: PMC7969712.	SVR, KRR, ENET, GBLUP, BayesB	Loblolly pine
18	Badji A, *et al*. Factors Influencing Genomic Prediction Accuracies of Tropical Maize Resistance to Fall Armyworm and Weevils. *Plants* (Basel). 2021 Jan; **10**(1):29. PMCID: PMC7823878.	Random forests	Maize
19	Maldonado C, *et al*. Genome-Wide Prediction of Complex Traits in Two Outcrossing Plant Species Through Deep Learning and Bayesian Regularized Neural Network. *Front Plant Sci*. 2020; **11**:593897. PMCID: PMC7728740.	Bayesian regularized neural network (BRNN), deep learning (long short-term memory), GBLUP, Bayseian regression	Zea mays L. and Eucalyptus globulus Labill
20	Pook T, *et al*. Using Local Convolutional Neural Networks for Genomic Prediction. *Front Genet*. 2020 Nov 12; **11**:561497. PMCID: PMC7689358.	LCNN	Maize
21	Medina CA, *et al*. Genome-Wide Association and Prediction of Traits Related to Salt Tolerance in Autotetraploid Alfalfa (Medicago sativa L.). *Int J Mol Sci*. 2020 May; **21**(9): 3361. PMCID: PMC7247575.	SVM, random forests	Alfalfa
22	Zingaretti LM, *et al*. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. *Front Plant Sci*. 2020; **11**: 25. PMCID: PMC7015897.	LCNN	Strawberry, blueberry
23	Azodi CB, *et al*. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. *G3* (Bethesda). 2019 Nov; **9**(11): 3691–3702. PMCID: PMC6829122.	Comparison of 6 linear and 6 non-linear algorithms, including: random forests, gradient tree boosting (GTB), ANN, and CNN	Maize, rice, sorghum, soybean, white spruce, switchgrass
24	Toda Y, *et al*. Predicting biomass of rice with intermediate traits: Modeling method combining crop growth models and genomic prediction models. *PLoS One*. 2020; **15**(6): e0233951. PMCID: PMC7304626.	Multiple regression, random forests	Rice
25	Yoosefzadeh-Najafabadi M, *et al*. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. *Front Plant Sci*. 2020; **11**:624273. PMCID: PMC7835636.	MLP, random forests, SVM	Soybean
26	Jin Y, *et al*. Advancing Agricultural Production With Machine Learning Analytics: Yield Determinants for California's Almond Orchards. Front Plant Sci. 2020; 11:290. PMCID: PMC7082403.	Random forests	Almond
27	Coulibali Z, *et al*. Site-specific machine learning predictive fertilization models for potato crops in Eastern Canada. *PLoS One*. 2020; **15**(8): e0230888. PMCID: PMC7413527.	Hierarchical Mitscherlich model, k-nearest neighbors, random forest, neural networks and Gaussian processes	Potato
28	Saffariha M, *et al*. Seed germination prediction of Salvia limbata under ecological stresses in protected areas: an artificial intelligence modeling approach. *BMC Ecol*. 2020; **20**:48. PMCID: PMC7456011.	Multiple linear regression, MLP	Salvia limbata
29	Garcia-Perez P, *et al*. Machine Learning Unmasked Nutritional Imbalances on the Medicinal Plant Bryophyllum sp. Cultured *in vitro*. *Front Plant Sci*. 2020; **11**:576177. PMCID: PMC7729169.	ANN	Bryophyllum sp
30	Hameg R, *et al*. Modeling and Optimizing Culture Medium Mineral Composition for in vitro Propagation of Actinidia arguta. *Front Plant Sci*. 2020; **11**:554905. PMCID: PMC7785940.	ANN	Actinidia arguta

Citation		AI Approach	Species
31	Hesami M, *et al*. Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. *Plant Methods*. 2020; **16**:112.	ANN, MLP, SVR	Chrysanthemum
32	Salehi M, *et al*. A hybrid model based on general regression neural network and fruit fly optimization algorithm for forecasting and optimizing paclitaxel biosynthesis in Corylus avellana cell culture. *Plant Methods*. 2021 Feb 5;**17**(1):13. PMCID: PMC7866739.	GRNN-FOA, MLP-GA	Corylus avellana
33	Hesami M, *et al*. Forecasting and optimizing Agrobacterium-mediated genetic transformation via ensemble model- fruit fly optimization algorithm: A data mining approach using chrysanthemum databases. *PLoS One*. 2020; **15**(9): e0239901. PMCID: PMC7526930.	MLP, adaptive neuro-fuzzy inference system (ANFIS), radial basis function (RBF)	Chrysanthemum

www.benchling.com