CellPress
REVIEWS

Review

Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence

Antoine L. Harfouche,[1,*] Daniel A. Jacobson,[2,3] David Kainer,[2] Jonathon C. Romero,[2] Antoine H. Harfouche,[4,5] Giuseppe Scarascia Mugnozza,[1] Menachem Moshelion,[6] Gerald A. Tuskan,[2] Joost J.B. Keurentjes,[7] and Arie Altman[6,*]

Breeding crops for high yield and superior adaptability to new and variable climates is imperative to ensure continued food security, biomass production, and ecosystem services. Advances in genomics and phenomics are delivering insights into the complex biological mechanisms that underlie plant functions in response to environmental perturbations. However, linking genotype to phenotype remains a huge challenge and is hampering the optimal application of high-throughput genomics and phenomics to advanced breeding. Critical to success is the need to assimilate large amounts of data into biologically meaningful interpretations. Here, we present the current state of genomics and field phenomics, explore emerging approaches and challenges for multiomics big data integration by means of next-generation (Next-Gen) artificial intelligence (AI), and propose a workable path to improvement.

Why Modern Plant Breeding Can Benefit from AI

Advances in breeding and agronomic practices for food crop improvement were largely responsible for the first green revolution, which doubled crop yields in less than 50 years [1,2]. If agricultural productivity is to be improved even more over the next 50 years, breeding must achieve unprecedented increases in yield and resource-use efficiencies while safeguarding harvests and preserving the environment and **ecosystem services** (see Glossary). Assessment based on a yield dataset comprising thousands of observations of wheat cultivars with diverse responses to weather conditions suggests that current breeding programs do not sufficiently prepare for climatic uncertainty and variability [3]. Consequently, the demand for climate resilience (i.e., the capacity to buffer against climate-related uncertainty and variability [4]) of crops must be better articulated [3]. The general goal of breeders is to make gains in yield by predicting which lines will produce the best progeny when crossed together. Advances in phenomics and genomics have provided unprecedented amounts of new data, which has allowed breeders to continue to push the yield trend upwards [5–8]. In particular, **genomic selection (GS)**, where the breeding value of an individual is predicted solely from genetic markers, has begun to supplant more traditional pedigree-based methods in several cereal and legume crops such as wheat, maize, soybean, and chickpea, as well as forest tree species such as eucalyptus, pines, and poplars [9–11]. Despite this success, the lack of predictive accuracy for many complex traits, such as yield, has revealed an inability to adequately **model** all of the relevant factors inherent to such traits. The process is complicated by the fact that the observed variation between individuals can be due to genetic (heritable) components, environmental components including farm management, and often an interaction between the two whereby an elite line may grow predictably in one environment but poorly in another [12].

Comparative and mapping studies have shown that while much of the observed variation is heritable, in many cases only a fraction of this heritability can be assigned to identified genetic factors such as SNPs or small insertions/deletions (indels). Several reasons may explain this 'missing heritability'. First, complex quantitative traits often appear to be governed by the infinitesimal or omnigenic model, in which many genes exert only a small effect and therefore go undetected unless very large populations are analyzed. Second, the relationship between the genotype and phenotype is not always linear and small changes on one hierarchical level may have a large impact on other levels. However, many statistical models fail to explore nonlinear relationships. Third, in biological systems genes are subject to complex regulatory networks, while their products often incur downstream modifications and interact with other pathways or protein complexes. Such conditional associations, known

Highlights

The integration of genomics and phenomics will speed the development of climate resilient crops; however, these omics technologies are generating large, heterogeneous, and complex data much faster than currently can be analyzed.

First-generation AI is being used in surveying and classifying omics data; however, it is designed to solve well-defined tasks of single-omics datasets that do not require integration of data across multiple modalities.

Next-generation AI can change the dynamics of how experiments are planned, thus enabling better data integration, analysis, and interpretation.

There is a critical need to develop means by which to open the black boxes prevalent in many current AI approaches so that they can be interpreted meaningfully from a complex biological perspective. AI decisions and outputs can be explained by breeders and researchers via human–computer interaction.

[1]Department for Innovation in Biological, Agro-food and Forest systems, University of Tuscia, Via S. Camillo de Lellis, Viterbo 01100, Italy

[2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[3]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN 37996, USA

[4]Unité de Formation et de Recherche en Sciences Économiques, Gestion, Mathématiques et Informatique, Université Paris Nanterre, 200 avenue de la République, 92001 Nanterre, France

as epistasis, are difficult to detect in studies that map genotype to phenotype with linear models due to low power and sheer computational demand.

The solution to the above problems begins with access to increasingly detailed and accurate data in more individuals, so that the complete underlying biological and environmental systems can be captured. Fortunately, the cost of genetic sequencing continues to plummet, innovative genomic assays are continually invented that may shed light on missing heritability and genetic regulation, and breeders have access to an ever-increasing suite of high-throughput sensors and imaging techniques for a wide range of traits and situations in the field. Epigenomics, transcriptomics, proteomics, metabolomics, phenomics, and **microbiomics** together with approaches to gather information about the microclimate and field environmental conditions have become routine. Omics technologies are aimed primarily at the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), metabolites (metabolomics), phenotypes (phenomics), and methylation profiles (epigenomics) in a specific plant sample. However, the ability to accurately predict and select the best lines, especially for specific environments, relies on our ability to model these immensely complex systems from the web of genomic and phenomic data at hand. Multiomics **big data** is a prominent example of such high-dimensional, heterogeneous datasets with very complex multilevel structures (Box 1). Increasingly, **AI** assists with this process (Box 1), as in other industries confronted with the challenge of big data. When utilized together, phenomics, genomics, and AI technologies can accelerate the development of climate resilient crop varieties with improved yield potential and stability and enhanced tolerance/resistance/resilience to anticipated and simultaneous abiotic and biotic environmental stresses, and deliver higher genetics gains in farmer's fields in less time (Figure 1, Key Figure).

Therefore, the primary goals of this review are to introduce the concepts and procedures of **Next-Gen AI**, to envisage how it can deal with these challenges and interface with multiomics big data to accelerate the breeding process for climate resilient crops, and to suggest future research directions.

Leveraging Next-Gen AI in Plant Breeding

AI has shown impressive results in fields such as image recognition [13,14] and has become a focus for big data analysis [15,16]. Current implementations of AI, such as **neural networks (NNs)** and extreme gradient boosting (XGboost) [17], have been focused nearly exclusively on predictive accuracy. In many cases, this accuracy comes at the cost of discernibility and explainability. Examples of these are NNs, which build nodes and paths that try to mimic brain neurons, and **deep learning (DL)** methods that incorporate multiple levels in a nonlinear hierarchical learner [18]. The inner workings and decision processes of these AI algorithms are opaque. Results can be seen but an understanding why a decision was made is lacking. Therefore, while AI has been a powerful tool for prediction and classification, it has not yet been a tool for **knowledge distillation**. The first step in the progression towards Next-Gen AI is the introduction of new **algorithms** of **explainable AI** that not only have a predictive model but also expose rules that are meaningful for human understanding (Box 2). This allows the researchers running the AI models to design better tests and obtain better data to improve future iterations.

Intrinsic to AI, multiple models or algorithms can be used to find the best fit for any given scenario, as stated by the 'no free lunch' theorem [19], which essentially establishes that there is no universally efficacious optimizer for all problems. That is, there is no guarantee that a model trained on one dataset will work on a different dataset [20]. Thus, it is important that each model be assessed for its appropriateness with respect to the problem. Likewise, the speed and efficiency of AI algorithms varies considerably, especially at the scale of big data. The recent introduction of the world's most powerful supercomputer, Summit, which has world's largest collection of graphical processing units (GPUs) (27 000), opens an exciting avenue for the integration of AI and scientific discovery. Therefore, Next-Gen AI will operate through multiple models and scales, compiling an **ensemble** [see Box 2 for the relevance to ensemble learning (EL)] of optimal learners for each subtask and taking advantage of ever-increasing computational power.

[5]Management and Strategy Department, EDHEC Business School, 59057 Roubaix, Cedex 1, France

[6]The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, Faculty of Agricultural, Food, and Environmental Quality Sciences, PO, Box 12, Rehovot 76100, Israel

[7]Laboratory of Genetics, Wageningen University & Research, 6708, PB, Wageningen, The Netherlands

*Correspondence:
aharfouche@unitus.it,
arie.altman@mail.huji.ac.il

With respect to plant breeding, Next-Gen AI is conceptually designed not only to predict breeding value for complex traits across environments and time scales, but to iteratively learn and improve. This requires the intelligent and efficient mining of data that truly represent the underlying systems biology and environment, plus interaction from humans at both the input and the output end. Breeders can thus better predict influences that affect yield and respond more quickly to changes not previously encountered. In addition, they can predict which variety or varieties would work best in a specific environment and what soil conditions are likely to be favorable. This open, collaborative approach will also lead to the utilization of underexploited germplasms, thus increasing the available genetic diversity in breeding gene banks. Next-Gen AI also has the potential to positively impact soil health and productivity by providing recommendations based on the metagenomics of microbial communities.

Field Phenomics Brings Opportunities for Accelerated Breeding

Observed phenotypic variation in living organisms is shaped by genomes, the environment, and their interactions [21]. Phenotyping plants under their natural and uncertain environmental conditions remains challenging due to the high level of **phenotypic plasticity** of many traits. However, plant breeding has led to a substantial reduction in the phenotypic plasticity of crops, probably due to the process of selection, canalizing many of the yield-related traits. Most of these traits are associated with regulation of plant water balance (e.g., stomatal conductance, transpiration, photosynthesis) and thus result in higher plant water uptake and lower water-use efficiency, leading to increased crop susceptibility to drought and other environmental stresses [22]. Hence, most empirical studies and current breeding efforts have been aimed towards improving abiotic stress tolerance. Unfortunately, the translation of phenotypic data generated from multiple genotypic sources and environments into practical **knowledge** remains limited. The integration of genotypic, environmental, and phenotypic data into meaningful knowledge reflecting the plant stress response profile is challenging, mainly due to the complexity of correlating the dynamic changes in environment to the phenotypic plasticity in a comparative way for many plants simultaneously. This challenge is known as the genotype–phenotype (GP) gap [23]. One of the ultimate goals of agronomic abiotic stress tolerance research is to identify the most relevant yield-related traits that are easy to measure as early as possible in the plant's life cycle, to enable the selection of the best-performing candidates for inclusion in further evaluations [24]. As many of these traits are highly (and almost instantaneously) regulated by the environment, screening should be done under as close as possible to natural-like unstable conditions; thus, the system must involve continuous monitoring of the plant environment (soil–atmosphere) as well as plant responses to changes in that environment. It is also important that the measurements be conducted simultaneously on all plants, as nonsimultaneous measurements may lead to the inadvertent comparison of traits under different ambient conditions.

One possible way to bridge the GP gap is the use of physiology-based gravimetric systems that enable direct measurement of the soil–plant–atmosphere-continuum (SPAC). This high-throughput functional phenotyping system (HFPS) enables the direct measurement and analysis of many physiological traits and their plasticity, hierarchy, and interactions [25].

Plant responses to the environment are dynamic both spatially (small differences at different locations of the growth zone) and temporally (from hourly changes in atmospheric conditions throughout the day to different climates throughout the growing season). Thus, phenotyping of complex traits not only is better represented under uncertain environmental conditions such as natural field conditions (or a good simulation of those conditions in controlled growth facilities), but also validates the authenticity of the measurements. Because of the feedback between plants and the environment, identical growth scenarios will be hard to repeat and analyze. For these reasons, deep phenotyping, involving continuous and simultaneous comparative measurements of many plants under multiple environmental conditions, coupled with Next-Gen AI will serve as a key tool to understand the plant–environment interactions and to unlock greater potential in data-to-knowledge transfer. Another important aspect that may derive from a better understanding of plant–environment interactions is the improvement of the genetic detection of stress response QTLs. High-throughput

Glossary

Algorithm: a set of well-defined computational instructions that extract, process, calculate, and estimate data to solve a problem.
Artificial intelligence (AI): a number of ML algorithms that build a model of rules learned from training data.
Big data: the digital convergence of structured data found inside databases and unstructured data flowing from new sources, such as sequencing platforms, sensors, satellites, and aerial- and ground-based platforms. This allows researchers to capture and analyze the data and make more informed decisions based on that information (Box 1).
Black box: a description of some ML systems (i.e., DL). They take an input and provide an output, but the calculations that occur in between are not easy for humans to interpret or understand.
Climatype: the climate and environmental variables found at the points of origin of evolved genotypes.
Data commons: a unified data repository that enables the plant research community to store, share, access, and interact with interoperable tools for analysis across multiomics studies in support of augmented breeding.
Data cube: a multidimensional dataset that generalizes matrices in a third dimension.
Data ecosystem: the infrastructure and applications that are used to capture, store, access, and analyze omics data. Sustaining the omics big data ecosystem requires coordinated efforts among stakeholders such as data creators, data maintainers, data users, and others.
Deep learning (DL) and reinforcement learning (RL): autonomous and self-teaching systems. DL takes raw features from an extremely large, annotated dataset (e.g., a collection of images or genomes) to train algorithms via various layers of artificial NNs to create a predictive tool based on patterns buried inside. Once trained, the algorithms can apply that training to analyze new data. RL dynamically learns by adjusting actions based on continuous feedback to maximize its performance.

phenotyping also has been adopted successfully to assess the genetics of estimated biomass dynamics in maize [26]. The integration of functional genomics with functional physiological phenotyping is also expected to yield an opportunity to better understand plant stress responses over time and in changing environments.

Field phenomics is a great suite of tools and methods to take on these challenges. Accelerated breeding for agriculturally relevant crop traits is key to the development of improved varieties and is critically dependent on high-resolution, high-throughput, field-scale phenotyping technologies that can efficiently discriminate among better performing breeding lines within a larger population and across multiple environments [27–30]. To be relevant to breeding programs, field phenomics must consider the nature of the environmental stress in the target environment and deploy multisensor unmanned aerial vehicle (UAV)- and ground-based platforms (Figure 1) that facilitate screening of thousands of field-grown genotypes and the development of more comprehensive data management, including crop modeling [29,31,32]. With the advent of novel sensors, high-resolution imagery, and new platforms for a wide range of traits, tissues, and conditions, phenomics has been championing the collection of more phenotypic data over the past decade [33–39]. Alongside genomics, phenomics is essential to support breeding programs and help breeders generate cultivars (commercial varieties) more adaptable to diverse and challenging environmental scenarios [40]. Using a combination of **genome-wide association studies (GWASs)** [41,42] and high-throughput phenotyping facilities, phenomics can serve as a novel tool for studies of plant genetics, genomics, gene characterization, and breeding [43].

Although data regularly captured by phenomics platforms, particularly plant imaging data, are being analyzed using a suite of **machine learning (ML)** methods [44], in the past few years DL has increasingly been used to make more sense of phenomics big data as it has a greater potential for further advancing image analysis [45]. For example, DL has greatly spurred the identification of plant features such as leaf counting [46,47], the derivation of vegetation indices from red–green–blue (RGB) images [48], the prediction of biomass traits [49,50], the detection of plant diseases [51], stress phenotyping [52,53], and the scoring of morphological and developmental phenotypes in genetic populations [54]. These DL approaches can be extended to any trait of interest to breeding for which high-resolution phenomics imaging data are available. Although these types of models typically operate as '**black boxes**' and require a leap of faith to believe their predictions, a recent study in soybean (*Glycine max*) sought to look under the hood of the trained model to explain each identification and classification decision made for a large class of stresses from RGB images of leaves [55].

Even if there is a good progress in image analysis and the spectral identification of phenomics data, profiling of plant dynamic physiological responses to changing environment status using these data is negligible. Namely, a single data point in time (i.e., an image taken above the canopy of the crop in the field) could be very different from the same image taken shortly after due to dynamic changes of soil–atmosphere conditions and the biological and physical responses. This is because of the multivariate and reciprocal effects of the collected data. Accordingly, a possible solution may be achieved by creating phenomics databases correlating the precise physiological profiles of plants under certain SPAC conditions to many phenomics methods. This detailed phenotypic data can be used as a basis for Next-Gen AI, enabling the structuring and fitting of the data into models that could be used for a better interpretation of field phenomic data.

As major breakthroughs in phenomics are witnessed [56], a long path to uncover and harness the complexities of integrating phenomic layers with genomic heritability and interactions with the environment still lies ahead, and ML is still far from fulfilling its potential in this research. Nonetheless, a great opportunity for Next-Gen AI is the ability to bridge that gap between phenomics and genomics.

New Frontiers in Genomics

Historically, the detailed high-throughput analysis of phenotypic responses to a variety of environments has taught us much about the physiological adaptation of plants to abiotic stresses. Applying

Ecosystem services: functions provided by the different taxa resident in complex ecosystems. The evolution of such co-dependencies allows communities of species to grow in environments in which any individual taxon may not be able to grow or thrive in isolation.
Ensemble: a set of individually trained algorithms whose combined outputs are more accurate than any of the single learning algorithms in the ensemble. EL combines the predictions of multiple trained NN models at once to achieve better generalization performance of learning algorithms and reduce the variance of predictions error (Box 2).
Explainable AI: the ability of algorithms to predict and provide interpretable explanations of their decisions. It allows researchers and human users to receive the best possible predictions and interpret the algorithm's outputs (Box 2).
Field phenomics: a field-based high-throughput phenotypic evaluation under the real conditions experienced by the plant.
Generative adversarial networks (GANs): deep NN architectures comprising two nets (generator and discriminator), pitting one against the other. They are trained in an unsupervised fashion to generate new data instances (the generator) and evaluate them for authenticity (the discriminator).
Genome-wide association studies (GWASs): scanning markers across the genomes of many individuals of a particular species to test for statistically significant associations between SNPs and phenotypes.
Genomic selection (GS): prediction models developed by estimating the combined effect of all existing markers simultaneously on a phenotype. Models are developed by phenotyping and genotyping a training population, so that all loci that regulate a phenotype are in linkage disequilibrium with at least one marker.
Knowledge: information or interpretations of the basic data (i.e., raw facts, observations) from a particular point of view, which has been validated and is thought to be true.
Knowledge distillation: a compression method for training a small model to mimic a

this knowledge to breeding programs aimed at developing resilient crops can provide a test case for Next-Gen AI. A Next-Gen AI breeding program requires detailed genomic data that can be accurately linked to adaptive traits. As a first step, modern breeding often starts with an assembly of a reference genome that is used as a basis for comparison between individuals of the species so that allelic variants can be identified, mapped, and eventually associated with phenotypic variation. Sequencing technologies have improved dramatically, allowing for more complete and accurate reference genome assembly. For example, the recently launched 10KP plan, as part of the Earth BioGenome Project [57], aims to sequence 10 000 different plant species including representatives of all known taxonomic families. For many species, this effort will provide a first genome-wide assembly, which may serve as a reference for subsequent resequencing initiatives that allow the detection of allelic variation within species [58]. This natural variation might then be exploited for the introgression of wild alleles in domesticated elite varieties, to improve the resilience of crops and forest trees to changing climate conditions. With each new genome sequence completed, opportunities expand for exploring genome regulation and variation and updating reference genome annotation using Ensembl browser release 93 [59]. Future strategies will also focus on the distribution of genomic data in a manner that enables researcher-driven analysis [59].

Parallel to the development of sequence technology, other omics profiling techniques have matured. RNA-seq has developed into a high-throughput method to quantify gene expression in multiple tissues at different developmental stages and under varied stresses [60]. ChIP-seq [61] and DNA affinity purification sequencing (DAP-seq) [60] reveal the genomic locations where transcription factors bind to DNA to perform their role as regulators of gene expression [60]. Further detail is provided by the assay for transposase-accessible chromatin sequencing (ATAC-seq) [62], which reveals the portions of chromosomes that are 'open chromatin' (i.e., accessible to regulatory binding [63,64]). Together these assays provide unprecedented insights into complex regulatory networks that often influence phenotype more than mutations in genes do. The metabolome and proteome can also now be accurately and comprehensively analyzed. The genetic analysis of these intermediates of the genotype-to-phenotype map, in relation to observed variation in the phenome, will greatly assist in elucidating the genetic architecture of complex traits and offer the opportunity to understand the flow of information that underlies plant responses to environmental stresses.

Our increasingly detailed access to systems biology data, from allelic markers and gene expression levels to expressed proteins, tissue-specific metabolite concentrations, and complex regulatory networks, gives us data layers that more closely reflect the true biological complexity underlying phenotypic variation. The challenge remains, however, of how to correctly integrate these data layers together, link them with environmental stress responses, and model the entire system accurately so that breeding can accelerate.

Linking the Genome to the Phenome: Next-Gen AI-Based GS

Current approaches to breeding climate resilient plants are focused on GS. The basic premise of GS is that the heritable (genetic) component of the trait can be viewed as having been generated by the combined effect of all underlying genome-wide variations (e.g., SNPs, indels), so models attempt to estimate the effects of each genetic variant on the phenotype while accounting for environmental effects. A statistical model is developed in a training population that has been both genotyped and phenotyped [65]. In essence, this is a task of mapping the genomic data layer (e.g., SNPs) to phenotypes (e.g., drought tolerance). The developed model is then applied to an independent breeding population that have only been genotyped, enabling prediction of their phenotype potentially years before the mature phenotype becomes measurable [65]. Higher accuracy of prediction can translate into increased yields in shorter breeding cycles.

If only genotypic and phenotypic layers of data are available to the model, the integration of regulatory circuits within and between omics layers is not taken into consideration. However, is well known that the variation of complex traits is subject to these regulatory circuits. Genomic signatures (e.g., alleles, haplotypes) that eventually affect the variation of a phenotype are often muted within the

pretrained model or ensemble of models that have been previously trained on large datasets. It is used to transfer the knowledge from the cumbersome model to a small model that is more suitable for deployment by minimizing the loss between itself and the larger model.
Machine learning (ML): the use of algorithms that can mathematically and statistically learn from data to extract important information, find hidden patterns, and make associations and predictions.
Microbiomics: the taxa and omics-based study of the microbiome - the collective omics of plant-associated microorganisms (i.e., bacteria, fungi, and microbes from other kingdoms).
Model: an equation that represents relationships and identifies patterns among features of a dataset.
Neural network (NN): a framework for many different brain-inspired ML algorithms to work together and process complex data inputs. It is a connectionist computational model, in which layers of neuron-like nodes mimic how human brains analyze information. The layers in a NN filter and sort information and communicate with each other, allowing each layer to refine the output from the previous one.
Next-generation (Next-Gen) AI: a framework that incorporates explainability, interpretability, EL (with a wide variety of learning architectures), and TL, uses the context of previously known information, and facilitates the use of human knowledge and experience for experimental/field trial design and results interpretation.
Phenotypic plasticity: the production of more than one phenotype from the same genotype when exposed to diverse environments.
Polytope: the generalized data space that extends 2D matrices into spaces of arbitrary dimensions. Omics and phenotypic measurements may each have *n* dimensions of environmental and temporal variables that form these complex polytopic spaces.
Transfer learning (TL): a ML technique in which an algorithm learns to perform one task and leverages that knowledge when learning a different but related task.

complexity of interacting omics layers. Higher omics levels, such as gene expression or metabolite concentrations, intrinsically integrate additive and epistatic signals from multiple genetic loci. It therefore stands to reason that omics layers can be better predictors of phenotype than SNPs alone due to their molecular proximity to the phenotype.

Currently, most GS models are limited to genotype and phenotype data. Therefore, mapping between genotype and phenotype is a highly challenging statistical task, since the biological distance between these layers obscures the true effect that SNPs have on the phenotype under varying conditions. Many approaches have been developed, including mixed-effect linear models, Bayesian models that endeavor to select only the most important predictive SNPs, and nonparametric ML and NNs [66]. Current GS models are prone to losing accuracy when the training population is not closely related to the breeding population (as might occur when predicting across different breeding cohorts) or when predicting across generations and/or environments. To date, NNs, which are superior predictors for many other specialized big data problems [18], have failed to consistently improve the accuracy of GS over statistical learning methods such as decision trees and simpler parametric models [66–69]. Why is this the case and what does the future hold for AI-based GS? Firstly, the strength of AI lies in its ability to find complex relationships and interactions within large datasets. However, since GS approaches today rely on mapping SNPs to phenotype, many other important data layers that explain trait variation have not been made available to the model. This gives an advantage to simpler models while depriving AI of its strengths. Secondly, GS models (including NNs) are typically treated as a black box. Data goes in, predictions come out, yet little is learned about the actual functional biology of the phenotype and it is therefore a struggle to improve the models iteratively.

From the breeders' perspective, if they have access to a rich set of omics and environmental data there are many ways in which they can manipulate the system to achieve a desired phenotype in the context of where the genotype will be deployed. By making changes to omics levels that lie between the genotype and the phenotype they can achieve a greater, and more refined, impact on the

Box 1. Making Sense of Big Data in the Era of Next-Gen AI

The term big data was coined by Cox and Ellsworth in 1997 and originally referred to data being too big to fit into memory and processed by conventional means [72,73]. Here, the definition of the eight Vs – volume, velocity, variety, variability, visibility, value, veracity, and vexing – is expanded (Figure I). Big data is now a reality in genomics and phenomics, with exciting opportunities arising from increased resolution and throughput in sequencing and phenotyping technologies. Without AI, big data would be overwhelming and chaotic, but by incorporating AI into big data analytics the technology becomes useful and rewarding. The combination of big data and AI has been referred to as both the fourth paradigm of science [74] and the fourth industrial revolution [75,76]. There are great opportunities at the intersection of big data and Next-Gen AI. However, there are equally great technical, scientific, and interpretive challenges to be tackled. Omics data present the raw material needed to gain insights into the complex biological mechanisms that underlie plant functions in response to environmental stresses. However, many datasets are noisy, sparse, and irregularly sampled or collected under different conditions and at ambiguous time points, resulting in ill-defined prediction targets. Furthermore, omics data are heterogeneous and high dimensional. They derive from a wide range of experiments that yield many types of information. The extent to which Next-Gen AI can help us solve complex biological questions depends heavily on the protocols, experimental settings, and standards in place for efficient metadata reporting and knowledgebase library development. Promising analytic solutions should align integration with research and incorporate prior knowledge into learning workflows. Big data also represents a powerful source for biologists to improve experimental design and research focus [77,78]. Applying a careful experimental design by placing controls randomly on experimental trials, using additional biological replicates, and capturing as much data as possible about environmental heterogeneity within and across field sites are important factors in generating good datasets with which to properly train an algorithm. Highly-dimensional big databases organized in many layers of data domains are imperative for omics data in order to capture consistent and high-quality phenotypic and genotypic information from various data sources. A clear research question alongside an appropriate computing infrastructure and robust statistical methods is required to extract important and relevant information that should guide follow-up experiments [79]. Future efforts that integrate broad, robust collections of phenotype and genotype data in combination with a greater understanding of data and relevant prior knowledge will create a rich resource for increasingly more efficient and detailed genome–phenome analysis to usher in new discoveries in plant breeding (e.g. [80]). Because data are only as good as the tools available to analyze them, the plant omics community must devise and develop specialized and publicly accessible data-management systems to reliably extract useful information from these data. An ideal data-management system would store data, provide common and secure access methods, and allow linking, annotation, and a way to query and retrieve information [81]. In addition, making multiomics data findable, available, identifiable, and reusable (FAIR) supports the reuse of the data and discoveries through good data management [82]. Taking advantage of international standards including the breeding API (BrAPI) also ensures integration and interoperability among several datasets [83]. Cloud and web computing that bring data and analysis together is crucial [84].

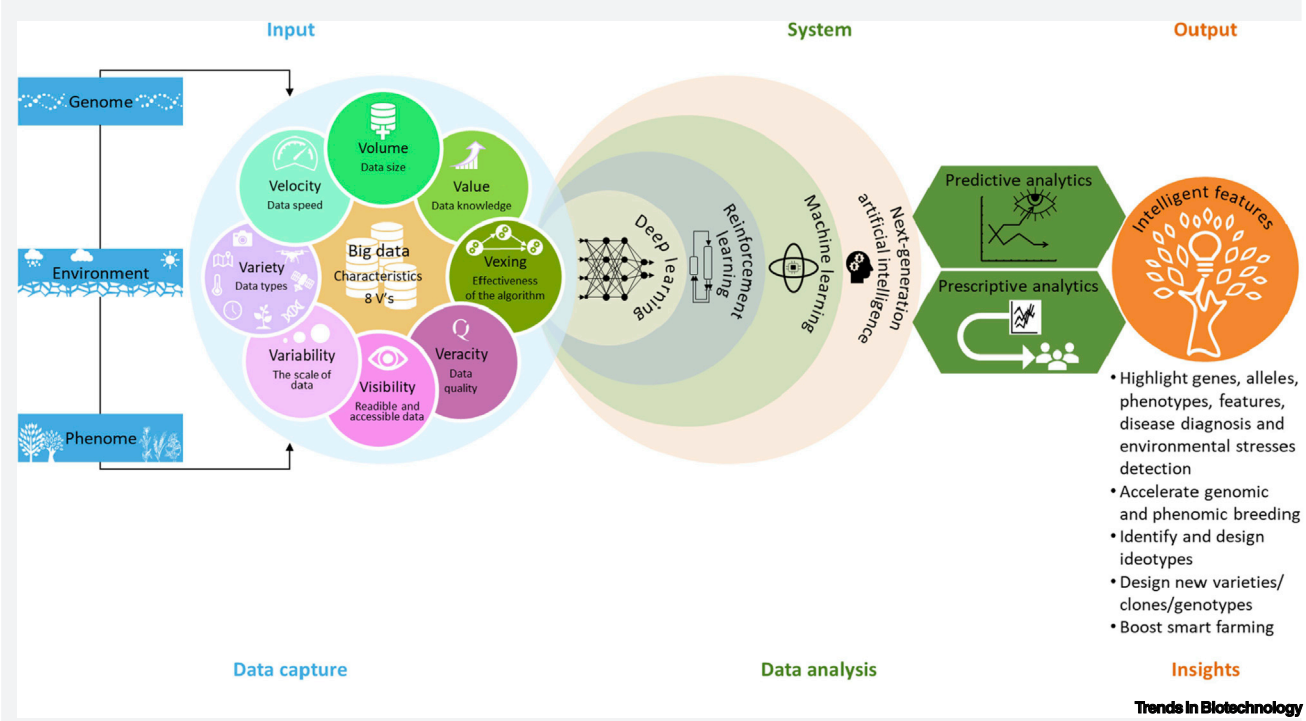

Figure I. The Combination of Big Data and Next-Generation (Next-Gen) Artificial Intelligence (AI) in Plant Breeding.

A schematic illustration of the 'input–system–output' big data analytics process of analyzing multiomics datasets and creating intelligent features in plant breeding innovation. The process comprises three major components – data capture, data analysis, and interpretation – that lead to insights on the underlying biology and optimized breeding choices. High-resolution measurements of genome, environment, and phenome through sequencing, sensing, and other high-throughput technologies are continuously increasing the amount of genotypic, environmental, and phenotypic data for large breeding populations for multiple traits in multienvironment field trials. There are eight important attributes of omics big data, known as the eight Vs: volume – the volume of image, sensor, and genome data; velocity – the speed at which data are ingested and processed; variety – the heterogeneity of data, including structured, semistructured, and unstructured; variability – data whose structure and meaning are rapidly changing; visibility – the visualization of data in a manner that is readable; veracity – the consistency, accuracy, and trustworthiness of data; value – extracting actionable insight or functional knowledge from data without loss of information; and vexing – the effectiveness of the modeling. Vexing is strictly related to the design of Next-Gen AI, which combines human knowledge and deep reinforcement learning for predictive and prescriptive analytics. Predictive analytics are concerned with estimating the likelihood of future outcomes based on statistics, modeling, and probabilities; it seeks to identify patterns in data and applies statistical models and algorithms to capture relationships between various datasets. Prescriptive analytics goes beyond predictive analytics by providing insights into the different possible actions to guide humans towards plausible solutions to complex problems; it quantifies the potential efforts of the different possibilities in order to advise on the best decision. This process can boost our analytic abilities to identify better alleles for key genes and superior lines and ideotypes for the development of climate resilient plants.

phenotype of interest. Specific alleles that affect critical gene expression and metabolite concentrations can be introgressed. Chemical hormones can be applied in the field and even the target environment is itself amenable to some manipulation (e.g., irrigation, fertilization). This potentially will allow models to hold for future generations and across populations, since the underlying biology is driving the accuracy rather than relatedness between individuals. However, it is important that models are capable of employing this data to predict which actions are optimal for the breeder's goals.

Next-Gen AI holds promise for GS, particularly if the omics data that are acquired capitalize on the strengths of AI and if explainable AI approaches are pursued that build on human knowledge to iteratively improve the model based on biologically validated outcomes. The acquisition of large-scale

Key Figure

Bridging Phenomics and Genomics: The Next Challenge

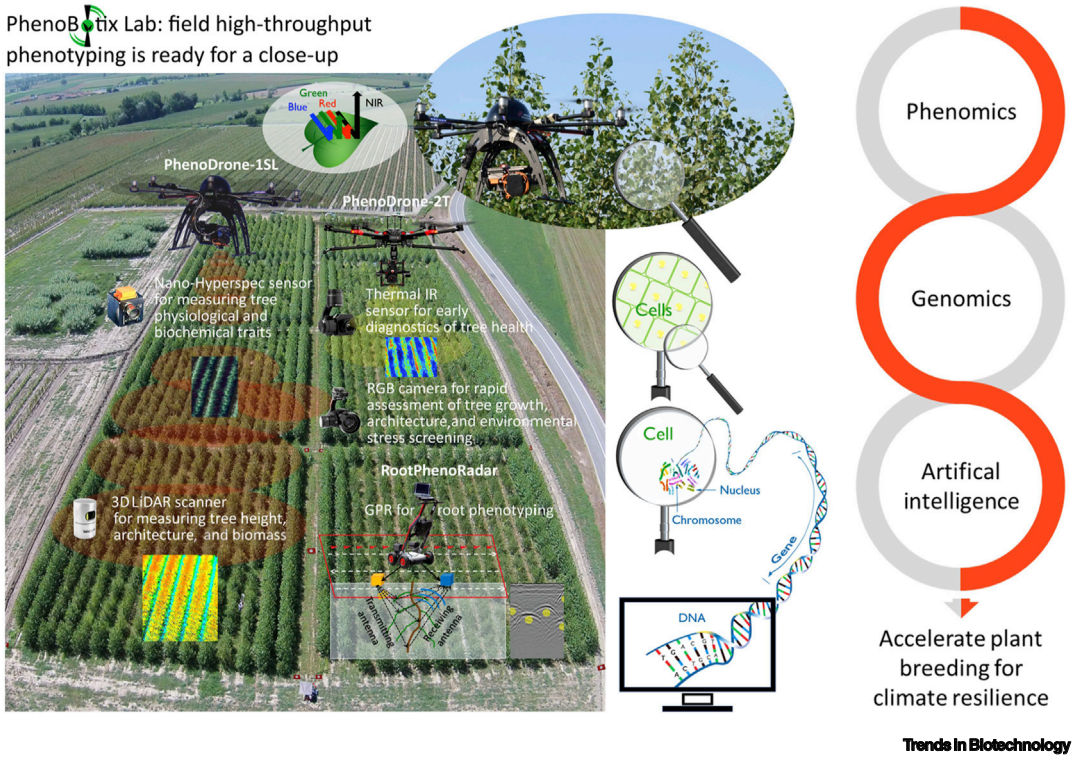

Figure 1. Next-generation phenomics, genomics, and artificial intelligence (AI) are perceived as key components of accelerating the plant breeding process. Platforms for field-based, high-throughput precision phenotyping of multiple quantitative traits are needed to complement the wealth of genomics information. For example, in the University of Tuscia PhenoBotix laboratory, unmanned aerial vehicles (UAVs), commonly referred to as drones, fly over breeding plots collecting data on individual plants and single leaves that computers will analyze and integrate with genomics data to make decisions about breeding. 'PhenoDrone-1SL' is equipped with integrated hyperspectral and light detection and ranging (LiDAR) sensors. Nano-Hyperspec is a high-spectral-resolution sensor that captures radiation reflected from plants in the visible–near-IR (VNIR) range of 400–1000 nm, which may contain information about leaf physiological status, water content, and biochemical traits in response to the environment. Micro-LiDAR is a 3D laser-based remote sensor that allows precise and consistent measurement of plant architecture, canopy height, and growth rates. 'PhenoDrone-2T' is equipped with thermal IR and red–green–blue (RGB) cameras. A thermal IR camera is a reliable and scalable phenotyping instrument for assessing canopy temperature and providing early diagnostics and quantification of plant responses to water stress. A RGB camera can be used to rapidly and objectively monitor stress response and formulate vegetation indices that provide information on plant health. Ground-penetrating radar (GPR) uses high-frequency radio waves and travels between plants collecting data on root architecture, which is vital to understand plant responses to drought stress and to breed crops with greater water-use efficiency and resilience in the face of severe environmental conditions. The collected streams of data from multiple types of sensors can be integrated using AI. AI creates an unprecedented opportunity for multiomics data analytics and knowledge discovery and will underpin efforts to develop plants with improved climate resilience. The photograph (left) is a drone aerial view of the poplar field trial of an F_2 breeding population in Savigliano, Italy aimed at dissecting the genetic architecture of drought tolerance in poplar. Phenotypic and genotypic screening of natural populations or germplasm collections is also of paramount importance for addressing the opportunities of AI in genotype–phenotype mapping.

phenomics and genomics data, in addition to the molecular layers between them, such as transcriptomics, proteomics, and metabolomics, will facilitate an era where AI models can find and explain complex interactions (Box 3); for example, predicting how changing water availability affects the expression of genes involved in plant growth while simultaneously impacting resistance to pests.

Next-Gen AI is intended to automate much of the analysis, but human input is also critical at multiple points in the process, which advocates for training and education of both computer scientists and biologists to fill the knowledge gaps. The first step forwards for breeding more climate resilient crops is to define the problem and the target space for the solution and to establish how to best take advantage of new algorithms, data collection, and computational power. This integration and definition of the target space is critical because, without direction, an algorithm is not guaranteed to produce a

Box 2. Algorithms, Explainable AI, Humans, and Communities

Here we highlight some emerging trends in Next-Gen AI that might maximize the impact of omics for plant breeding.

More Knowledge from Smaller Datasets

Algorithms require large, well-annotated, and clearly labeled datasets deriving from a variety of experimental, environmental, and physiological stress conditions so that they can efficiently learn to distinguish features and categorize patterns (Table I). To circumvent this requirement, researchers need to become better at making the data (or metadata) associated with their studies accessible in a machine-readable format. Ground-truth data can also be exceptionally valuable. Another promising solution is metalearning, whereby knowledge is learned within and across problems [85,86]. In addition, TL or pretraining – the ability of an algorithm to improve learning capacities on one given small dataset (target) through previous exposure to a large dataset (source) – becomes more applicable [87]. For example, TL allowed NN algorithms to apply image classification prowess acquired from one data type, rodent cells, to another type, human cells [88]. TL also helps to reduce overfitting as the model generalizes well from training data to unseen data. Recently, a new deep TL (DTL) approach was reported to make significant progress in extracting information from complex biomedical images [89]. Furthermore, due to the fact that each of the standard single learning algorithms has its own advantages and disadvantages (in terms of bias and learning performance), EL, such as bagging (parallel ensemble method) and advanced bagging, has been increasingly used to learn an ensemble of classifiers for collaborative classification. This approach compensates for the disadvantages of individual classifiers and improves the overall accuracy of classification [90]. By combining bagging and boosting (parallel and sequential ensemble approaches), the prediction accuracy and speed of a broad range of applications, and under a variety of scenarios, have been significantly improved [91]. Another option is to use **generative adversarial networks (GANs)** to generate *in silico* data with properties of real data. For example, the *Arabidopsis* rosette image generator AN (ARIGAN) was used to generate synthetic rosette-shaped plants [92]. This can be extended to multiomics datasets; for example, to generate larger gene expression datasets that can be exploited to build predictive models of transcriptional regulation [93].

Explainable AI

The black box nature of AI models remains a great challenge for genomics and phenomics applications. The lack of explainability is endemic to most black-box models and represents a major barrier to wider use [94]. This challenge underscores the importance of explainable AI that attempts to overcome these limitations. Explainable AI systems are able to both reason and explain their behavior and decisions to researchers and human users. As many of these systems are opaque in their operations, new approaches are available to provide highly accurate and meaningful explanations with no loss of prediction accuracy [95–98]. In plant breeding, explainable AI is urgently needed for many purposes including phenomics and genomics. Plant stress phenotyping remains predominately a tedious and time-consuming manual rating exercise that is mainly based on visual symptoms performed by trained plant scientists. An interesting example is provided by a recent study in which a robust and accurate explainable DL model has been successfully used to not only automate the process of plant stress identification, classification, and quantification [99] but also to explain which visual symptoms are used to make predictions [55]. Ongoing and emerging developments in the application of explainable AI approaches in genomics and phenomics will enable us to envision an exciting future for plant breeding.

Plant Breeding Research Produced When Researchers and Communities Work Together Is Better for Society

Next-Gen AI platforms must be built within the context of the problem that researchers and breeders are solving and in collaboration with farmers and AI industry experts. Both cognitive systems and end users must be trained together as part of a symbiotic relationship. AI companies and plant breeding and biotechnology companies must be prepared to invest in training technology users as much as they are training the system itself. Knowledge generated in partnership with seed banking should also be encouraged. Phenomics, genomics, and Next-Gen AI enable the screening and analysis of large amounts of germplasm and offer a major expansion in the genetic variation available for breeding.

The Promise of Education and Knowledge Transfer

Lack of education of and knowledge transfer to breeders and farmers is often suggested as a significant bottleneck. To effectively transfer knowledge to practice, community research networks and infrastructures such as the EU Plant Phenotyping Network (EPPN2020)[i], the EU Infrastructure for Multiscale Plant Phenotyping and Simulation for Food Security in a Changing Climate (EMPHASIS)[i], EU Cooperation in Science and Technology (COST) Action FA1306[iii] on the Quest for Tolerant Varieties: Phenotyping at Plant and Cellular Level, the International Plant Phenotyping Network (IPPN)[iv], the Australian Plant Phenotyping Facility (APPF)[v], and the North American Plant Phenotyping Network (NAPPN)[v] as well as crop-specific initiatives such as the Wheat Initiative of the Group of Twenty (G20)[vii] and the Consultative Group on International Agricultural Research (CGIAR)[viii] centers in the Excellence in Breeding Platform are playing a major role in generating critical mass and stimulating interactions between researchers, breeders, and farmers. All of these initiatives will enable greater technology uptake by breeders and farmers and integrate the community globally.

Stage		Step	Approach	Platform/Algorithm		
1	Data capture	Data sources	Environment and climate monitoring	Platform[b]	Satellite and drone remote sensing, weather stations, microclimates, geographic information systems (GIS), geospatial, soil monitoring stations	
			Phenomics		Satellites, UAVs, field scanning, phenotyping towers, autonomous ground vehicles/rovers/tractors, proximal sensing carts, glasshouses	
			Genomics, transcriptomics, proteomics, metabolomics		Next-generation sequencing (NGS), short-read technologies, long-read technologies, Hi-C technology, gas/liquid chromatography–mass spectrometry	
		Data collection	Extraction, transformation, and loading (ETL)[c]		[d]Vertical distribution of data integration	
					[e]Horizontal distribution of data integration	
2	Data storage	Store and stream 'in-motion' semistructured and unstructured data - nonrelational (NoSQL) databases	Data lakes for storing data as objects and associated metadata		Distributed file systems	
		Store, index, and query 'at-rest' structured data - structured query language (SQL) databases	Multidimensional **data cubes**		Data warehouses	
3	Data preprocessing	Dimensionality reduction	Feature extraction	Algorithms[f]	Principle component analysis (PCA), t-stochastic neighbor embedding (t-SNE), partial least square (PLS)	
			Feature selection		Least absolute shrinkage and selection operator (LASSO), elastic net, ridge regression, recursive feature elimination–support vector machine (RFE-SVM), correlation-based feature selection (CFS)	
4	Data analysis	Segmentation	Clustering		Markov clustering, spectral clustering, association rules, independent component analysis	NNs, iterative random forest (iRF), Markov chain Monte Carlo (MCMC), Q-learning, deep Q-network (DQN), Bayes optimal classifier, deep autoencoder, deep belief networks (DBNs), deep forest, GANs[g]
			Classification		Decision trees (DTs), support vector machine (SVM), random forests (RFs), naïve Bayes classifier, logistic regression, nearest neighbor classifier, latent Dirichlet allocation (LDA)	
		Prediction	Regression		Linear regression (LR), Markov chain, regression trees, temporal difference (TD)	
5	Data interpretation	Explanation	Feature importance/ model selection		Local interpretable model-agnostic explanations (LIMEs), Shapley additive explanations (SHAPs), DL important features (DeepLift), integrated gradient, random intersection trees (RITs)	

Table I. Representative ML, DL, Reinforcement Learning (RL), and EL Algorithms and Platforms across the Data-Capture-to-Interpretation Chain[a]

[a]Technology infrastructure such as servers, networking, virtual machine, operation systems, middleware, and runtime can be on premises or in the cloud.

Table I. *Continued*

[b]Various phenomics platforms use a range of sensors and cameras including, but not limited to, RGB, thermal IR, light detection and ranging (LiDAR), multispectral, hyperspectral, fluorescence, ground-penetrating radar, and electromagnetic inductance.
[c]Extraction involves retrieving data from their sources then sourcing the data required for analytics; transformation involves cleansing and harmonizing data (e.g., filtering, removing duplicates) from their sources to the target through a declarative, traceable, reusable data pipeline; loading involves the movement of data to data stores, streams, and analytics tools.
[d]Vertical distribution refers to the partitioning of the different omic layers in multitiered architectures across a cluster of computers.
[e]Horizontal distribution deals with the distribution of a single omics layer across multiple computers.
[f]All of these algorithms are relevant for ML, DL, RL, and EL. Because the field of artificial intelligence is developing rapidly, it will give rise to new algorithms and many variants in the future.
[g]These algorithms combine two or more deep NNs or classes of DL and **RL** algorithms. Many of these algorithms may fall into multiple categories.

model that is predictive of the target space that was intended. A simple example of this is building a model with too broad a solution space (i.e., the whole planet) to predict the best place to grow a crop, resulting in a solution that the best place for said crop is on land. This information may be technically accurate, but it is not useful. The same can be said about the data used as input for a model. A human, for example, can be expected to predict the effects of heavy rainfall only if he or she has been exposed to previous occasions of rainfall or gained knowledge about rainfall from others. Likewise, an organism with improved drought tolerance can be bred only if the genotypic data for that organism can be linked to the phenotype (i.e., its responses to drought conditions). The more complete and accurate the data and the more relevant the target space, the better the inferences; it is then up to the breeder to ensure that the algorithmic output is relevant to the problem.

All AI algorithmic models have a starting point that significantly influences the rate of learning and the accuracy of the final model. The more informed the starting point, the more accurate the final model is expected to be and the faster that model will converge. In humans, this is analogous to being quicker to learn new information in a subject you are already knowledgeable about than to learn new information in a field you have no prior experience in. In ML, the process of preconfiguring the starting parameters is called a 'warm start' and generally requires human insight to initiate appropriately. Another process, **transfer learning (TL)**, describes giving a model a warm start by applying information learned from another previously trained model. Judging whether such a transfer of knowledge from one model to another is appropriate is another instance that requires a human decision. In the field of biology, this often entails the detailed study of a model organism and the application of the knowledge gained from that system to other organisms. This must be done in an intelligent way; knowledge from *Arabidopsis* is likely to transfer more informatively to alfalfa (*Medicago sativa*) than to cattle (*Bos taurus*). These types of insights come naturally to humans and should represent the type of conclusion that Next-Gen AI can come to as well. From a practical standpoint, it is important for Next-Gen AI to have access to the genomics and phenomics of multiple species beyond the species of immediate interest so that TL becomes an option. More data is always better for the model as long as the Next-Gen AI can appropriately weight the information according to its relevance.

Although much data can be gathered in an automated manner, Next-Gen AI will require the knowledge and rationality of humans (e.g., researchers, breeders, farmers) to evaluate the outcomes, because for any given scientific question there are multiple approaches that can result in a solution. There are many different algorithms that can address a dataset and problem from alternative perspectives, resulting in additional insights and levels of confidence. In the case of ensembling, advantage can be taken of the no free lunch theorem and the use of multiple algorithms to produce multiple solution spaces. The intersection of these solution spaces indicates agreement between different methods and represents a 'best-way' forwards approach. The total solution space allows multiple algorithms to address various edge cases that one algorithm may excel at handling in comparison with others. Each of these algorithmic solutions is a line of evidence that can be used in conjunction with each other and with external knowledge of the breeder to determine the best course of action. Possibly the most important input humans will have in Next-Gen AI is in the case of external validation, including specific testing of identified features. Models can be predictive within given training and validation sets and can provide hypotheses on the functioning of an organism. However, external

Box 3. Analyzing Polytopic Spaces: A Key Challenge for Next-Gen AI

Historically, AI methods have focused on the use of an X matrix (e.g., an image) that is to be regressed against a Y matrix or vector (e.g., an image label). While this has been very powerful it is a simplistic view of a complex biological system. It would be desirable to be able to include multiomics layers as well as temporal, geospatial, and environmental variables in the model and as such the representation becomes a series of **polytopes** of arbitrary dimension. However, given a set of polytopic data structures, how can one find combinatorial patterns within and across those structures that represent a biological organism and its interaction with its environment? At present there are no available AI methods that can handle this level of complexity. An algorithm that is capable of both building an accurate prediction from multiple data layers to multiple data layers and finding the combinatoric interactive elements within and between those layers epitomizes the goal for Next-Gen AI (Figure I). Such an algorithm would be able to use the data collected from each of the omic layers plus environmental data combined with priors taken from validated research to produce a model to derive previously unknown, important biological interactions. The most applicable classes of algorithm for this task are likely to be a DL approach such as convolutional NNs (CNNs), or decision tree-based approaches like iRF.

CNNs can introduce various types of omics data at different node layers in the NN architecture, thus more appropriately representing underlying biological relationships than most other statistical models. Where NNs struggle, however, is to explain the underlying biological significance that drives the outputs of the model. iRF, on the other hand, intrinsically has a very interpretable structure, but is limited to explaining how one data layer explains one feature of interest at a time. Currently in development, however, is an expansion of iRF known as tensor iRF (TiRF), which is designed to model across the multiple polytopic space and may provide one of the first tractable Next-Gen AI solutions for systems biology (Figure I). For example, using TiRF one would be able to use SNPs, gene expression, and environmental data measured across time and sites in a set of genotypes to predict the phenomic layer as well as find sets of genes and environmental variables that affect each of the phenotypes and combinations thereof.

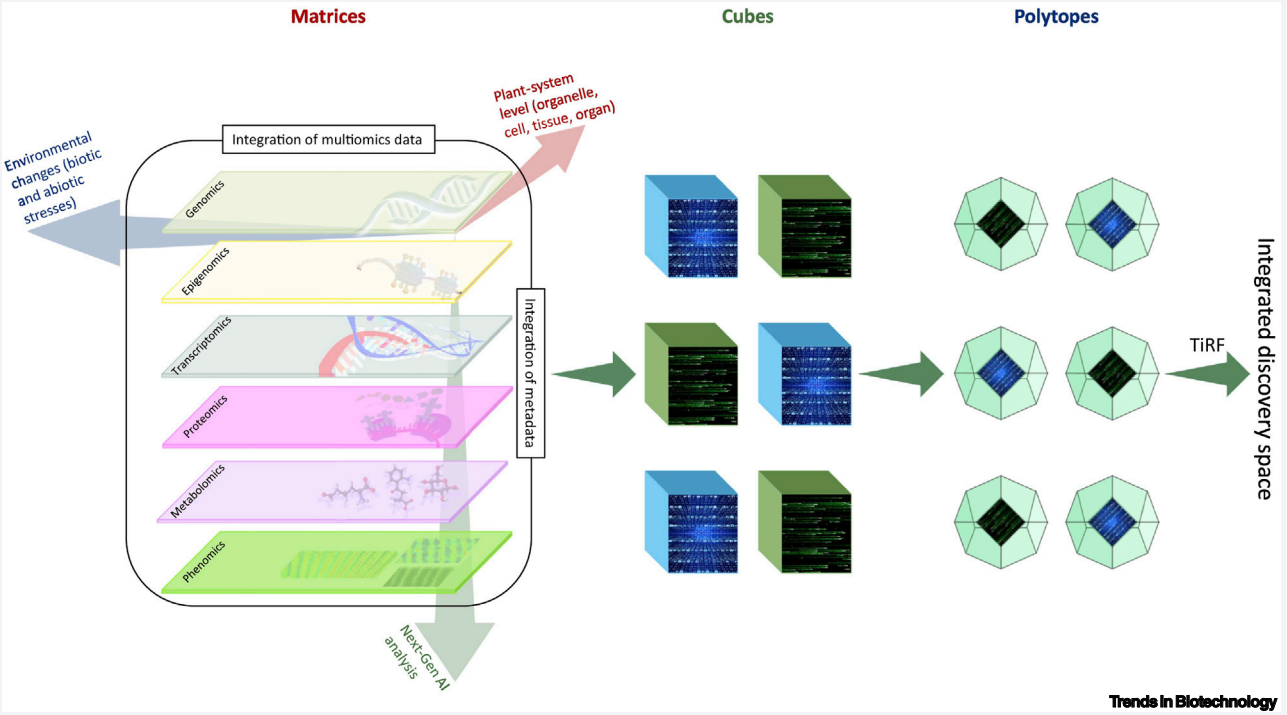

Figure I. Finding Combinatorial Effects in Multiomics Big Data: Towards the Development of More Sophisticated Artificial Intelligence (AI) Algorithms.

Extending the use of omics and systems biology approaches is necessary to understand complex biological processes through the integration of datasets (morphological, physiological, molecular, biochemical) at the level of a defined system (plant organelle, cell, tissue, organ; see red arrow). Notably, these components are under the influence of environmental changes (biotic and abiotic stresses; see blue arrow). The integration of multiomics data, including mutations defined through genomics, methylation profiles through epigenomics, mRNA levels through transcriptomics, protein abundance and type through proteomics, metabolite levels through metabolomics, genotype and environment contribution to phenotypic variations through phenomics, and metadata will enable us to create a global picture with higher informative power than single omics data. Ultimately, the combination of multiomics data with temporal, spatial, and **climatype** data yields polytopes of arbitrary dimensions that are beyond the capabilities of current AI models. The vertical green arrow highlights the next-generation (Next-Gen) AI models that enable the analysis of an organism in a dynamic multiomics fashion (genomics, epigenomics, transcriptomics, proteomics, metabolomics, phenomics). The development of more sophisticated AI algorithms, such as tensor iterative random forest (TiRF), will allow new integrated discovery spaces to be created that will have a large impact on breeding.

validation is required, including field, wetlab, or greenhouse testing, to verify predictive accuracy and incorporate new biological insights from the results of these tests into the next set of AI models. This feedback loop of information distillation, hypothesis generation, and knowledge refinement can continue iteratively to accelerate crop breeding and improvement. This includes the potential optimization of crop yield across different environments and conditions.

Looking forwards, breeders should be encouraged to capture as much data as possible beyond genotype and phenotype, especially as costs for obtaining such data decrease. Armed with this data, Next-Gen AI, and a specific goal in mind, breeders can determine the mechanisms that they can best employ to reach their optimal phenotypes and adaptations to targeted environments.

The Path Upwards in Next-Gen AI: From Augmented Breeding to Smarter Farming

Modern breeders are able to gather ever-growing amounts of data, and with the proposed Next-Gen AI they will be able to do more with that big data than ever before (Box 4) to support sustainable agriculture. It is impossible for any human to truly take advantage of all of this data to connect between the data layers or to understand what should be implemented in practice to optimize yield or resilience. Consider, however, that breeders are conceptually NNs with decades of focused training plus thousands of years of accumulated knowledge of breeding, climate, and biology. These breeders can inform the inputs to the Next-Gen AI models defining boundaries on practicality and economic goals. The more integrated breeders are in this process, the more they can improve the models; the better the models, the better the outcomes (Figure 2). As breeders are able to incorporate changes suggested by Next-Gen AI, they will be able to test the improvements that the model predicts, but by the next growing/breeding cycle the model will have become even more refined by the new data. The more breeders are integrated into the **data ecosystem**, the more data the models will have (Figure 2). Consequently, plant breeding research coproduced when researchers and communities work together is more likely to be useful to society (see Box 2 for the relevance coproduction).

Moreover, genomic resources can also be interrogated to identify targets for increased productivity under different environmental conditions, focusing particularly on the interactions between genetics, environment, and management (G×E×M) on trait plasticity [70]. In this case, M, which is basically a way to modify E, can simply be included in prediction models as an environmental variable. M may

Box 4. An Example Application of the Agricultural Use of Next-Gen AI

UAVs (a.k.a. 'drones') take time-series phenotypic data, flying daily over the target crop and using a powerful hyperspectral snapshot camera to accurately capture hundreds of wavelengths of light. The light striking each pixel is broken down into many different spectral bands to provide significant morphological, physiological, and biochemical information on what is imaged. This information is passed through a NN that has been trained to describe this hyperspectral imagery as phenotypes of the crop, such as height, chlorophyll content, and flowering. The NN is initially trained by a human (farmer, breeder, or researcher) manually labeling a dataset with the appropriate phenotype values, effectively transferring the farmer's knowledge to the NN. After this training, the predictive process is automated by the NN. The information revealed by the NN is then combined with known data, such as other sensors on the plot that measure environmental conditions, and with the known genotypic data of the plants. Using this combined dataset, if the human needs to find the watering pattern that promotes the best growth, he or she would apply a suite of algorithms in Next-Gen AI (see Table I in Box 2), such as a second NN, to predict the ideal amount of water and inform the human of the optimal watering schedule. Alternatively, if the human needs to select the next generation of the crop, another ML algorithm, such as iterative random forest (iRF) [100], would use the plethora of phenotype values stored in the database to predict the parents to cross that will produce the best set of progeny. It is the human's prerogative to determine his or her end goals and desired traits. As this is an explainable AI model, it would be possible to understand each prediction made, as the model would report the sets of interacting features that it is using for the prediction. The availability of such an explainable AI model would have significant implications in farming and plant breeding as the model could be deployed in mobile platforms or as a mobile application.

lead to a significant impact on the ability of farmers to obtain valuable information from their fields and thus to better control them during the growing season.

Wide-scale integration between field and algorithm is already underway for GS in cattle breeding. GS was pioneered in cattle, and the cattle breeding industry continues to maintain the cutting edge of this field [71]. Multiple cattle farmers upload periodically measured phenotypes to a centralized database, which continually augments the pedigree and data records, followed by reiteration of the GS algorithm to inform selection choices. It stands to reason that plant breeding will follow suit in the future. Barriers to integration, such as wireless data speeds and automation, are rapidly falling away with fourth-generation (4G) and fifth-generation (5G) mobile networks and augmented global positioning systems (GPSs) that provide increased accuracy and integrity of GPS information to farmers.

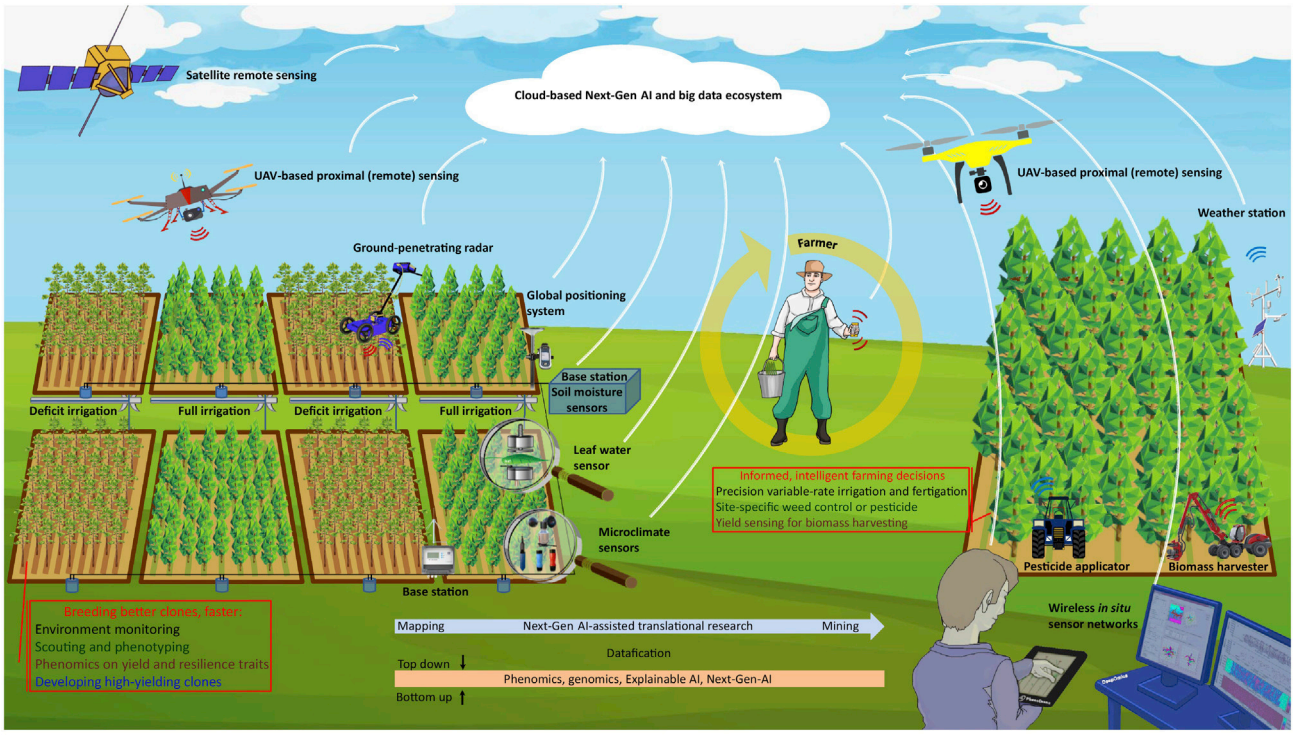

Figure 2. Next-Generation (Next-Gen) Artificial Intelligence (AI)-Augmented Farm: Reimagining the Farm of the Future.

Figure360▷ For a Figure360 author presentation of Figure 2, see the figure legend at https://doi.org/10.1016/j.tibtech.2019.05.007. Field phenomics offers high-throughput, nondestructive technologies to quantify plant performance in response to the environment. Sensors mounted on platforms such as satellites, drones, and farm machinery can be used to rapidly and precisely measure, over time and on high numbers of individuals, relevant plant attributes in the field and inform selection decisions. In parallel, time-series microclimate data such as ambient air temperature, humidity, solar radiation, and soil moisture level can also be repeatedly collected with ground-based sensor networks. The resulting large phenotype and environmental datasets will be linked to genomics data collected on individual plants and used as a reference on which data analysis pipelines can be developed. A cloud-based omics big data platform in plant breeding provides a single large dataset and computing ecosystem that can store, analyze, and share these data. Next-Gen AI-enabled algorithms can then be used to evaluate breeding decisions and predict which variety or varieties will show the best performance in field testing. These algorithms will ultimately accelerate the breeding process and enable breeders to scale their best ideas to the size of their breeding pipelines. With advances in computing power, algorithms can be rapidly retrained with additional data to overcome the environmental variability challenge and improve their predictive capability. Next-Gen AI will require farmers to actively participate in the development and rapid deployment and adoption of superior food and tree crops. Farmers will also benefit from direct on-farm applications of Next-Gen AI to predict crop yield and detect drought, pests, and diseases with high spatial resolution. Combining phenomics, genomics, big data, and Next-Gen AI can potentially help farmers make informed and intelligent planting, farming, and management decisions. This data-driven systems approach, where multidisciplinary and multisector collaborations are instrumental, will help to create successful digital and computational climate-smart agriculture and forestry solutions for the future.

Concluding Remarks and Future Directions

It is crucial that plant breeding education adapts to the digital revolution. Researchers and breeders must become adept at weighing machine-generated advice against farmers' needs. Generating knowledge for plant breeding is of limited value unless researchers also have the capacity to transform such knowledge into practice. Needed are knowledge–action approaches that incorporate additional skills and perspectives that will help to produce knowledge used for the achievement of augmented breeding and smarter farming.

Agriculture will rely on Next-Gen AI methods that make decisions and recommendations from big data that are representative of the environment and a systems biology-based understanding of a plant. Next-Gen AI is envisioned to enable breeding to perform at higher levels than previously possible, efficiently utilizing highly heterogeneous and complex data.

One of the important challenges in the decade ahead will be the ability of individual researchers and breeders to submit and share their own data to a **data commons**. Through this approach, they may gain a deeper understanding of their own data, have greater control over how the data is being used by the research community, and make large, or expensive-to-collect, datasets available to all. This will be an important aspect of breeder–researcher partnered breeding efforts and will improve the robustness of breeding programs. Without data sharing, it would be all but impossible for a single research group to screen thousands of breeding lines and obtain sufficient data to support the kinds of studies that Next-Gen AI requires.

AI-based approaches are now common workloads in cloud data centers. Thus, cloud data centers are increasingly providing large numbers of GPUs to support AI applications. NVIDIA (the leading manufacturer of GPUs) has recently determined that AI workloads are generating the majority of network traffic in cloud data centers and has purchased Mellanox (the leading manufacturer of adaptive routing InfiniBand backplanes which connects computer nodes within data centers) as a strategic effort to ensure that there is adequate bandwidth in GPU-heavy data centers. Furthermore, wireless wide area network connectivity is rapidly expanding, and telecommunications companies worldwide are making significant investments in next-generation wireless network technologies (e.g., 4G, 5G) and other companies are investing heavily in satellite deployment to provide global internet access. Thus, high-bandwidth network connectivity will become increasingly ubiquitous, even in remote areas. These industry trends would seem to indicate that cloud-based infrastructure for AI applications will be increasingly readily available and therefore farmers and breeders will be able to load data into cloud-based AI applications from handheld, drone, or farming machinery platforms.

Impressive phenomics and genomics results using ML and DL have been reported. As encouraging as these results are, they are not good enough yet to contemplate complete reliance on the technology to speed breeding, which remains largely a demanding, time-consuming, and costly task. Regardless of improvements in the efficiency of data generation, the plant research community still struggles when stepping into the translational processes. Genomics, epigenomics, transcriptomics, proteomics, metabolomics, and phenomics are still mainly separate fields that generate limited knowledge when viewed in isolation. Multiomics data should be used and integrated concomitantly to accelerate the plant breeding process.

The good news is that the plant research community is producing a stunning array of solutions to most of these challenges. However, to be considered a success such solutions must have the potential to be scaled up and adopted widely. The central challenge will therefore not be a knowledge gap but translation failure. Already established research infrastructures must continue to intersect in meaningful ways. Where possible, capacity building and technology transfer components should be integrated into these efforts. There is also an urgent need for concerted action to create a framework for research and development that coordinates and finances AI innovations in plant breeding. Public and private venture capital and partnerships between the two should consider the potential effect of

increasing their own investments in AI-augmented plant breeding research and innovation. With proper management and support, these investments can yield valuable returns in terms of discoveries and socioeconomic impact.

As Next-Gen AI becomes routine and prominent, our focus will necessarily shift from the technical performance of algorithms to the new models of farming that hopefully enable a new agricultural revolution that is better for both people and the environment.

Outstanding Questions

When will phenomics and genomics technologies increase their throughput significantly and thus reach their full potential for plant breeding?

Agricultural environments are uncontrolled and unpredictable; how arduous will Next-Gen AI testing and validation tasks in a farmer's field be?

When will Next-Gen AI infrastructure, analytics, and applications be available for a wide community of researchers and companies focused on plant breeding?

What are the largest remaining obstacles for small-scale Next-Gen AI implementation? How might computer processing power, in-house analytics capabilities, and/or cloud-based data ecosystems, in which data are deposited in a large pool of computational resources, be built and used to analyze massive amounts of image and sequence data at multiple scales?

If the implementation of longer-term plant breeding AI projects is expected, in what ways might the academic, breeding, farming, and AI communities be encouraged to interact, facilitate interdisciplinary research, and boost the two-way transfer of knowledge?

How can public, private, and venture capital investments be attracted for proof-of-concept Next-Gen AI pilot projects? Is it more useful for entrepreneurial firms to look at Next-Gen AI through the lens of business capabilities and high potential business value rather than technologies?

Acknowledgments

Funding was provided by the EU 7th Framework Programme – WATBIO, grant no 311929 (A.L.H. and J.J.B.K.), the Italian Ministry of Education, the University & Research Brain Gain Professorship to A.L.H., and the Center for Bioenergy Innovation, a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Funding was also provided by the DOE, Laboratory Directed Research and Development funding (ORNL AI Initiative ProjectID 9613) at the Oak Ridge National Laboratory. The iRF-driven GS work referred to in this review used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. The authors would like to acknowledge Ashley Cliff for manuscript review and editing. The manuscript was coauthored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Resources

[i]https://eppn2020.plant-phenotyping.eu/

[ii]https://emphasis.plant-phenotyping.eu/

[iii]www.plant-phenotyping.org/home_costfa1306

[iv]www.plant-phenotyping.org/

[v]www.plantphenomics.org.au/

[vi]http://nappn.plant-phenotyping.org/

[vii]www.wheatinitiative.org/

[viii]http://excellenceinbreeding.org/

References

1. Duvick, D.N. (2005) The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* 86, 83–145
2. Lopes, M.S. *et al.* (2012) Genetic yield gains and changes in associated traits of CIMMYT spring bread wheat in a "historic" set representing 30 years of breeding. *Crop Sci.* 52, 1123–1131
3. Kahiluoto, H. *et al.* (2019) Decline in climate resilience of European wheat. *Proc. Natl. Acad. Sci. U. S. A.* 116, 123–128
4. Carpenter, S. *et al.* (2001) From metaphor to measurement: resilience of what to what? *Ecosystems* 4, 765–781
5. Kole, C. (2013) *Genomics and Breeding for Climate-Resilient Crops*, Springer
6. Crain, J. *et al.* (2018) Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *Plant Genome* 11, 170043
7. Brown, T.B. *et al.* (2014) TraitCapture: genomic and environment modelling of plant phenomic data. *Curr. Opin. Plant Biol.* 18, 73–79
8. Clifton-Brown, J. *et al.* (2019) Breeding progress and preparedness for mass-scale deployment of perennial lignocellulosic biomass crops

switchgrass, miscanthus, willow and poplar. *GCB Bioenergy* 11, 118–151
9. Wallace, J.G. *et al.* (2018) On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* 52, 421–444
10. Harfouche, A. *et al.* (2012) Accelerating the domestication of forest trees in a changing world. *Trends Plant Sci.* 17, 64–72
11. Crossa, J. *et al.* (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975
12. Voss-Fels, K.P. *et al.* (2018) Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* 132, 669–686
13. Cordero-Maldonado, M.L. *et al.* (2019) Deep learning image recognition enables efficient genome editing in zebrafish by automated injections. *PLoS One* 14, e0202377
14. Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv*, 1409.1556.
15. Obermeyer, Z. and Emanuel, E.J. (2016) Predicting the future – big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375, 1216–1219
16. Njah, H. *et al.* (2019) Deep Bayesian network architecture for big data mining. *Concurr. Comput. Pract. Exp.* 31, e4418
17. Chen, T. and Guestrin, C. (2016) XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, pp. 785–794, ACM
18. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
19. Kawaguchi, K. *et al.* (2017) Generalization in deep learning. *arXiv*, 1710.05468.
20. Wolpert, D.H. and Macready, W.G. (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67–82
21. Li, X. *et al.* (2018) Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6679–6684
22. Dalal, A. *et al.* (2017) To produce or to survive: how plastic is your crop stress physiology? *Front. Plant Sci.* 8, 2067
23. Gosa, S.C. *et al.* (2018) Quantitative and comparative analysis of whole-plant performance for functional physiological traits phenotyping: new tools to support pre-breeding and plant stress physiology studies. *Plant Sci.* 282, 49–59
24. Moshelion, M. and Altman, A. (2015) Current challenges and future perspectives of plant and agricultural biotechnology. *Trends Biotechnol.* 33, 337–342
25. Negin, B. and Moshelion, M. (2017) The advantages of functional phenotyping in pre-field screening for drought-tolerant crops. *Funct. Plant Biol.* 44, 107
26. Muraya, M.M. *et al.* (2017) Genetic variation of growth dynamics in maize (*Zea mays* L.) revealed through automated non-invasive phenotyping. *Plant J.* 89, 366–380
27. Furbank, R.T. and Tester, M. (2011) Phenomics – technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644
28. Fahlgren, N. *et al.* (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr. Opin. Plant Biol.* 24, 93–99
29. Araus, J.L. *et al.* (2018) Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23, 451–466
30. Shakoor, N. *et al.* (2017) High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol.* 38, 184–192
31. Ludovisi, R. *et al.* (2017) UAV-based thermal imaging for high-throughput field phenotyping of black poplar response to drought. *Front. Plant Sci.* 8, 1681
32. Dungey, H.S. *et al.* (2018) Phenotyping whole forests will help to track genetic performance. *Trends Plant Sci.* 23, 854–864
33. Atkinson, J.A. *et al.* (2019) Uncovering the hidden half of plants using new advances in root phenotyping. *Curr. Opin. Biotechnol.* 55, 1–8
34. Goggin, F.L. *et al.* (2015) Applying high-throughput phenotyping to plant–insect interactions: picturing more resistant crops. *Curr. Opin. Insect Sci.* 9, 69–76
35. Kyratzis, A.C. *et al.* (2017) Assessment of vegetation indices derived by UAV imagery for durum wheat phenotyping under a water limited and heat stressed Mediterranean environment. *Front. Plant Sci.* 8, 1114
36. Han, L. *et al.* (2018) Clustering field-based maize phenotyping of plant-height growth and canopy spectral dynamics using a UAV remote-sensing approach. *Front. Plant Sci.* 9, 1638
37. Díaz-Varela, R.A. *et al.* (2015) High-resolution airborne UAV imagery to assess olive tree crown parameters using 3D photo reconstruction: application in breeding trials. *Remote Sens.* 7, 4213–4232
38. Liebisch, F. *et al.* (2015) Remote, aerial phenotyping of maize traits with a mobile multi-sensor approach. *Plant Methods* 11, 9
39. Watanabe, K. *et al.* (2017) High-throughput phenotyping of sorghum plant height using an unmanned aerial vehicle and its application to genomic prediction modeling. *Front. Plant Sci.* 8, 421
40. Camargo, A.V. and Lobos, G.A. (2016) Latin America: a development pole for phenomics. *Front. Plant Sci.* 7, 1729
41. Kooke, R. *et al.* (2016) Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in *Arabidopsis*. *Plant Physiol.* 170, 2187–2203
42. Fusari, C.M. *et al.* (2017) Genome-wide association mapping reveals that specific and pleiotropic regulatory mechanisms fine-tune central metabolism and growth in *Arabidopsis*. *Plant Cell* 29, 2349–2373
43. Yang, W. *et al.* (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5, 5087
44. Tsaftaris, S.A. *et al.* (2016) Machine learning for plant phenotyping needs image processing. *Trends Plant Sci.* 21, 989–991
45. Pound, M.P. *et al.* (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* 6, gix083
46. Ubbens, J. *et al.* (2018) The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant Methods* 14, 6
47. Giuffrida, M. *et al.* (2018) Pheno-Deep Counter: a unified and versatile deep learning architecture for leaf counting. *Plant J.* 96, 880–890
48. Khan, Z. *et al.* (2018) Estimation of vegetation indices for high-throughput phenotyping of wheat using aerial imaging. *Plant Methods* 14, 20
49. Chen, D. *et al.* (2018) Predicting plant biomass accumulation from image-derived parameters. *Gigascience* 7, giy001
50. Aich, S. *et al.* (2017) Deepwheat: estimating phenotypic traits from images of crops using deep learning. *arXiv*, 1710.00241.

51. Mohanty, S.P. *et al.* (2016) Using deep learning for image-based plant disease detection. *Front. Plant Sci.* 7, 1419
52. Mahlein, A.-K. *et al.* (2018) Hyperspectral sensors and imaging technologies in phytopathology: state of the art. *Annu. Rev. Phytopathol.* 56, 535–558
53. Singh, A.K. *et al.* (2018) Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci.* 23, 883–898
54. Wang, X. *et al.* (2019) High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. *bioRxiv*. Published online January 23, 2019. https://doi.org/10.1101/527911.
55. Ghosal, S. *et al.* (2018) An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4613–4618
56. Tardieu, F. *et al.* (2017) Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, R770–R783
57. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
58. Varshney, R.K. *et al.* (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530
59. Cunningham, F. *et al.* (2018) Ensembl 2019. *Nucleic Acids Res.* 47, D745–D751
60. Boyle, A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322
61. Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* 316, 1497–1502
62. Buenrostro, J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213
63. Sherwood, R.I. *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* 32, 171
64. Raj, A. *et al.* (2015) msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS One* 10, e0138030
65. Hayes, B.J. *et al.* (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443
66. Heslot, N. *et al.* (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146
67. Bellot, P. *et al.* (2018) Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819
68. Montesinos-López, A. *et al.* (2018) Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 (Bethesda)* 8, 3813–3828
69. Montesinos-López, O.A. *et al.* (2018) A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3 (Bethesda)* 9, 601–618
70. Boyles, R.E. *et al.* (2019) Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *Plant J.* 97, 19–39
71. Weigel, K.A. *et al.* (2017) A 100-year review: methods and impact of genetic selection in dairy cattle – from daughter–dam comparisons to deep learning algorithms. *J. Dairy Sci.* 100, 10234–10250
72. Cox, M. and Ellsworth, D. (1997) Application-controlled demand paging for out-of-core visualization. In *Vis '97. Proceedings of the 8th Conference on Visualization*, pp. 235–244, IEEE Computer Society Press
73. Cox, M. and Ellsworth, D. (1997) Managing big data for scientific visualization. *ACM Siggraph* 97, 21–38
74. Hey, T. *et al.* (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research
75. Gil, Y. *et al.* (2014) Amplify scientific discovery with artificial intelligence. *Science* 346, 171–172
76. Schwab, K. (2017) *The Fourth Industrial Revolution*, Crown Business
77. Dolinski, K. and Troyanskaya, O.G. (2015) Implications of big data for cell biology. *Mol. Biol. Cell* 26, 2575–2578
78. Decker, S.R. *et al.* (2018) High throughput screening technologies in biomass characterization. *Front. Energy Res.* 6, 120
79. Angerer, P. *et al.* (2017) Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* 4, 85–91
80. Cobb, J.N. *et al.* (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.* 126, 867–887
81. Boyle, J. (2013) Biology must develop its own big-data systems. *Nature* 499, 7
82. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018
83. Selby, P. *et al.* (2019) BrAPI – an application programming interface for plant breeding applications. *Bioinformatics*. Published online March 23, 2019. https://doi.org/10.1093/bioinformatics/btz190.
84. Marx, V. (2013) The big challenges of big data. *Nature* 498, 255
85. Jankowski, N. *et al.* (2011) *Meta-Learning in Computational Intelligence*, Springer
86. Smith-Miles, K.A. (2009) Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Comput. Surv.* 41, 6
87. Yosinski, J. *et al.* (2014) How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* (Jordan, M.I. *et al.* eds), pp. 3320–3328, MIT Press
88. Webb, S. (2018) Deep learning for biology. *Nature* 554, 555–557
89. Kim, S.-J. *et al.* (2018) Deep transfer learning-based hologram classification for molecular diagnostics. *Sci. Rep.* 8, 17003
90. Liu, H. and Cocea, M. (2018) Nature-inspired framework of ensemble learning for collaborative classification in granular computing context. *Granul. Comput.* Published online July 30, 2018. https://doi.org/10.1007/s41066-018-0122-5.
91. Arsov, N. *et al.* (2017) Generating highly accurate prediction hypotheses through collaborative ensemble learning. *Sci. Rep.* 7, 44649
92. Valerio Giuffrida, M. *et al.* (2017) ARIGAN: synthetic *Arabidopsis* plants using generative adversarial network. In *Proceedings of the 2017 Computer Vision Problems in Plant Phenotyping, 2017 International Conference on Computer Vision Workshops (ICCVW)*, pp. 22–29, IEEE
93. Camacho, D.M. *et al.* (2018) Next-generation machine learning for biological networks. *Cell* 173, 1581–1592

94. Castelvecchi, D. (2016) Can we open the black box of AI? *Nature* 538, 20–23
95. Štrumbelj, E. and Kononenko, I. (2014) Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* 41, 647–665
96. Ribeiro, M.T. *et al.* (2016) Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM
97. Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Jordan, M.I. *et al.* eds), pp. 4765–4774, MIT Press
98. Lundberg, S.M. *et al.* (2018) Consistent individualized feature attribution for tree ensembles. *arXiv*, 1802.03888.
99. Singh, A. *et al.* (2016) Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21, 110–124
100. Basu, S. *et al.* (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1943–1948