CellPress
Partner Journal

Molecular Plant
Perspective

Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction

Yunbi Xu[1,2,3,*], Xingping Zhang[3], Huihui Li[1,6], Hongjian Zheng[4], Jianan Zhang[5], Michael S. Olsen[7], Rajeev K. Varshney[8], Boddupalli M. Prasanna[7] and Qian Qian[1]

[1]Institute of Crop Sciences, CIMMYT-China, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2]CIMMYT-China Tropical Maize Research Center, School of Food Science and Engineering, Foshan University, Foshan, Guangdong 528231, China

[3]Peking University Institute of Advanced Agricultural Sciences, Weifang, Shandong 261325, China

[4]CIMMYT-China Specialty Maize Research Center, Shanghai Academy of Agricultural Sciences, Shanghai 201400, China

[5]MolBreeding Biotechnology Co., Ltd., Shijiazhuang, Hebei 050035, China

[6]National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya, Hainan 572024, China

[7]CIMMYT (International Maize and Wheat Improvement Center), ICRAF Campus, United Nations Avenue, Nairobi, Kenya

[8]State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, Australia

***Correspondence: Yunbi Xu (y.xu@cgiar.org)**

https://doi.org/10.1016/j.molp.2022.09.001

ABSTRACT

The first paradigm of plant breeding involves direct selection-based phenotypic observation, followed by predictive breeding using statistical models for quantitative traits constructed based on genetic experimental design and, more recently, by incorporation of molecular marker genotypes. However, plant performance or phenotype (P) is determined by the combined effects of genotype (G), envirotype (E), and genotype by environment interaction (GEI). Phenotypes can be predicted more precisely by training a model using data collected from multiple sources, including spatiotemporal omics (genomics, phenomics, and enviromics across time and space). Integration of 3D information profiles (G-P-E), each with multidimensionality, provides predictive breeding with both tremendous opportunities and great challenges. Here, we first review innovative technologies for predictive breeding. We then evaluate multidimensional information profiles that can be integrated with a predictive breeding strategy, particularly envirotypic data, which have largely been neglected in data collection and are nearly untouched in model construction. We propose a smart breeding scheme, integrated genomic-enviromic prediction (iGEP), as an extension of genomic prediction, using integrated multiomics information, big data technology, and artificial intelligence (mainly focused on machine and deep learning). We discuss how to implement iGEP, including spatiotemporal models, environmental indices, factorial and spatiotemporal structure of plant breeding data, and cross-species prediction. A strategy is then proposed for prediction-based crop redesign at both the macro (individual, population, and species) and micro (gene, metabolism, and network) scales. Finally, we provide perspectives on translating smart breeding into genetic gain through integrative breeding platforms and open-source breeding initiatives. We call for coordinated efforts in smart breeding through iGEP, institutional partnerships, and innovative technological support.

Key words: smart breeding, genomic selection, integrated genomic-enviromic selection, spatiotemporal omics, crop design, machine and deep learning, big data, artificial intelligence

Xu Y., Zhang X., Li H., Zheng H., Zhang J., Olsen M.S., Varshney R.K., Prasanna B.M., and Qian Q. (2022). Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol. Plant. **15**, 1664–1695.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

INTRODUCTION

Climate change, human population growth, and arable land loss have driven the field of plant breeding to seek innovative approaches to produce improved varieties for more sustainable crop production (Xiong et al., 2022). The phenotype (P) or performance of a plant is the result of the genotype (G), growth environment (E), and genotype by environment interactions (GEIs). Many indirect selection indices have been developed to efficiently select genotypes based on quantitative genetics for phenotypic selection (PS) of target traits and their components (Baker, 1986). In many cases, the phenotypic performance or target trait can be predicted using genetic models constructed with phenotypic and genotypic information. Various genetic models, including the best linear unbiased prediction (BLUP) procedure, have been developed for prediction (Bernardo, 2021).

Classical breeding pipelines can be improved by incorporating new data-centric technologies for increasing genetic gain in breeding programs (Wallace et al., 2018; Crossa et al., 2021). With the development of molecular markers, genetic variation across the whole genome can be effectively captured. Thus, marker-trait association can be established and used to develop genetic models to predict phenotypes (e.g., Lande and Thompson, 1990). As a general method, genomic selection (GS) was developed to predict complex traits from genotypic data based on a model constructed using a training population that has been genotyped and phenotyped (Meuwissen et al., 2001; Bernardo and Yu 2007; Heffner et al., 2009).

In addition to the molecular marker and phenotype information that have long been used in predictive breeding, multiomics information, involving genomics, phenomics, enviromics, epigenomics, transcriptomics, metabolomics, microbiomics, and metagenomics, are being generated with increasing complexity. However, many of these data types have seldom been used in predictive breeding. Full utilization of "omics" requires a strong capacity to handle multidimensional big data, integrate information from multiple sources, retrieve useful genetic and molecular information through comparative analysis and artificial intelligence (AI), and develop optimized genetic models.

To better understand crop phenotypes, environmental factors need to be fully explored. A large volume of envirotypic data has been amassed in breeding activities, for example, through multi-environmental trials (METs). However, enviromic data have seldom been utilized in predictive breeding and instead have historically been considered as a single collective component (E) in interpreting GEIs and all the unexplainable variations (Flores et al., 1998; Kang, 2002; Xu, 2010). The E component should be treated more comparably with G and P for improved predictive breeding.

Bringing fully informative G, P, and E into the simple but universal formula P = G + E + GEI will generate more powerful genomic prediction (GP) models. However, the translation of G-P-E data into meaningful knowledge remains a genotype–phenotype gap (Gosa et al., 2018). Recently, multiomic (including enviromic) data have been used in predictive breeding in several ways, by developing predictive models (Costa-Neto et al., 2021a; Cooper and Messina, 2021), enviromic similarity (Costa-Neto et al., 2021b), and the environmental index (EI) (Li et al., 2018, 2021b; Guo et al., 2020) to enable integrated modeling and prediction, and by defining climatic or landscape-based variables as enviromic markers (Resende et al., 2021). Plant breeding is expected to become smarter in the near future with the integration of more data types and increasingly sophisticated and improved predictive models.

AI can increase the probability of identifying truly favorable genotypes by focusing on current breeding materials with the potential to achieve optimal traits. Multinational seed enterprises (MSEs) have been using AI or AI-like approaches for predictive breeding of major crops for more than a decade. With the support of stakeholders and funding agencies, AI will continue to evolve and create new opportunities for plant breeding (Lee, 2021). This article will address challenges in the field of predictive breeding and develop a potential smart breeding strategy. This strategy incorporates an integrated genomic-enviromic prediction (iGEP) or selection by combining multiomics information with innovative breeding technologies that are driven by big data and AI.

BIG DATA AND MULTIDIMENSIONAL BREEDING INFORMATION

Plant breeding has been evaluated by the genetic gain that can be achieved annually. For PS and GS, we have the following formula for evaluation of genetic gain: $\Delta G = ih^2\sigma_p/t$ and $\Delta G = ir_A\sigma_A/t$, respectively, where i is the standardized selection differential, σ_p is the SD of population trait values (phenotypic basis), r_A is the correlation between genomic estimated breeding value and true breeding value, σ_A is the additive variance, and t is time (in units appropriate for breeding-program cycles, e.g., years). The goal of iGEP is to improve r_A by developing and optimizing predictive models using all available big data and AI technologies. Predictive breeding helps reduce the cycle time required to "solve" the breeding puzzle, producing a plant with the desired combination of traits, adapted to specific and changing environments.

The history of plant breeding can be divided into four existing or near-future stages (Wallace et al., 2018). The first stage largely consists of incidental selection performed by farmers. The second stage involves experimental design and statistical analysis to improve selection effort, with the involvement of phenotypic prediction. Marker-assisted breeding, including GP, has been incorporated at the third stage. We are able to combine all favorable alleles/haplotypes into optimal combinations at the fourth stage. Smart breeding will enable effective use of both traditional and big data generated in the past as well as in the present and future.

A new era for breeding with big data

The phrase "big data" refers to datasets that are too large to fit into local memory and cannot therefore be processed by conventional means (Cox and Ellsworth, 1997). More generally, big data has been defined as "extensive datasets"—primarily in terms of volume, velocity, and/or variability (NIST, 2015). In comparison with smaller datasets, the sheer volume of big data can help distinguish the part from the whole, the local from the global, the current from the historical, and thus, "the tree from the

forest." It also helps us to distinguish signal from noise, cause from association, and essence from phenomenon. We can make decisions through comparisons of big data in many ways, for example, at micro and macro scales, from finite and infinite sources, and with zoom-out and zoom-in images.

Breeding-related big data have some specific and important attributes, referred to as "the 9 Vs" (Figure 1). Plant breeding programs often produce enormous datasets from myriad sources, most of which are structured and can be organized in Excel sheets or databases. However, a great deal of relevant unstructured data are also produced, such as emails, social media posts, data- and word-processing documents, web pages, and audio, video, and photo files. Sorting and extracting value from these unstructured data are more difficult. Smart breeding programs generate and use both structured and unstructured data (Table 1).

Big data technologies: Data processing, data mining, and cloud computing

Modern breeding programs face the challenges of collecting, archiving, and mining big data from multiple sources. The raw data generated in plant breeding must be processed into a format that can be used for computer-based analysis (e.g., machine learning [ML]) (Figure 2). Data processing includes cleaning (identifying and replacing incomplete, inaccurate, irrelevant, or otherwise problematic data), integration (combining data from different sources and formats), transformation (performing normalization or concept hierarchy to suitably format the data), reduction (presenting a less complex representation of the data, e.g., through dimensionality reduction [DR], numerosity reduction, and data compression), discretization (replacing continuous raw values with interval ranges), and sampling (selecting a subset of samples from which to estimate the characteristics of the entire population). After processing, data can be mined through statistical analyses, ML, artificial neural networks (ANNs), or pattern discovery. Some of these methods are used in modeling and prediction.

As we have scaled our breeding and begun to collect data spatiotemporally, it has become virtually impossible to sift through the ocean of data to perform predictive breeding without the assistance of innovative technologies. Big data technologies, services, and tools, such as Hadoop, MapReduce, Hive, and NoSQL/NewSQL databases, and data integration techniques, in-memory approaches, and cloud technologies, have emerged to help meet the challenges posed by the flood of web, social media, internet of things, and machine-to-machine data. Bringing big data under the cloud roof presents great opportunities and advantages, allowing researchers to focus on data analysis and mining instead of managing servers and databases.

Cloud computing can be used to deliver a full set of computing services over the internet (Figure 2). Cloud-based scalable environments make it possible to deploy applications for data on the scale of zettabytes. Cloud computing also simplifies connectivity and collaboration within and between organizations, providing access to relevant analytics and streaming data sharing. Many enterprises, including MSEs, have adopted cloud computing to improve their IT operations and develop better software more quickly. The major challenges of big-data cloud implementations include network dependency, latency issues, and decreased control over security and compliance (Chan, 2018).

"Tri-typing" technologies and multidimensional breeding information

The plant breeding pipeline begins with natural and artificial populations and ends with the release of commercial varieties or novel germplasm. In between, many types of big data are collected: empirical breeding, selection indices, parental and cross combinations, combining ability, hybrid performance, parental relationships (relativeness and pedigrees), genetic distances, developmental and growth records, dynamic variation, environmental factors, and varietal replacement or renewal (Figure 2). Most of the data collected fall into three dimensions (G-P-E). Collecting these data spatiotemporally introduces additional dimensions.

Phenotyping

We discuss phenotyping first because tri-typing began with the collection of phenotypic data in classical field trials and precision and high-throughput phenotyping (HTP) becomes a bottleneck compared with other typing technologies. Phenotyping can be performed at various and ever-finer scales, such as populations, individuals, plots, fields, plant parts, tissues, cells, molecules, and metabolism. However, several phenomics issues must be seriously considered: (1) scientific issues involving the whole phenome (integrated phenotypes), relationships with other omics, genetics of phenotypes, and environmental effects; (2) technical issues including throughput, precision, cost, automation, integration with other data, and data management; and (3) the phenomic industry, such as agronomy and crop production.

Phenotyping has evolved from capture by visual observation and simple facilities to HTP supported by improved facilities (Araus et al., 2018). With the advent of HTP technologies, including novel sensors and high-resolution imagery, an underutilized reservoir of data has been generated over the past decade (Kyratzis et al., 2017; Watanabe et al., 2017; Araus et al., 2018; Atkinson et al., 2019; Harfouche et al., 2019; Pieruschka and Schurr, 2019; Watt et al., 2020; Jin et al., 2021). HTP solutions in plant breeding can be complicated by many factors that influence output data: research goals, applications, environments, crop species, plant parts and tissues, and growth media. Two major HTP methods are generally used: plant-to-sensor and sensor-to-plant (Li et al., 2021a). The former uses fixed sensors, and the plants are brought to the sensors for detailed phenotyping. This option may not be feasible if it is impractical or impossible to move the studied plant, and the phenotyping environment may vary significantly from the field conditions under which the plant grows, limiting the utility of the collected data. The alternative is the sensor-to-plant method, in which sensors are moved close to the plant. This can be achieved through the use of gantries, spidercams, tractors, or unmanned aerial vehicles.

High-throughput, nondestructive field phenomics can be used to quantify plant performance in specific environments. Compared with their wild ancestors, modern crops are often planted in genetically uniform stands with high densities and improved characteristics, such as steeper leaf and root angles and shorter

Breeding Events	Program design • Breeding platforms • Germplasm resources • Genetic variation • Prediction and selection • Generation acceleration • Multi-environmental trials • Plant variety protection
Datafication	Activities • Conversations • Words • Excel sheets • Numbers • Voice • Photos • Videos • Sensors • Social media • Browser logs • Networks

Properties	Definition	Conventional breeding	Smart breeding
Volume	The size of dataset to be analyzed and processed	Kilobytes to gigabytes, largely comprising pedigrees, phenotypes, METs and genotypes, with relatively small sample sizes	Frequently larger than terabytes or petabytes, comprising G-P-E from multiple sources collected spatiotemporally, with large sample sizes and many populations/breeding lines
Velocity	The speed at which data are generated, produced, created, or refreshed (updated)	Low speed, up to several megabytes per day, largely through manual observation and facility-supported processes with low- to medium-throughput	Continuous and massive flow of data (several megabytes or more per day) from various omics and G-P-E across time and space, supported by high-throughput facilities with low costs
Variety	The heterogeneity of data, including structured, semi-structured, and unstructured	Largely structured with defined formats and lengths, such as G & P data for germplasm, populations and breeding lines	In addition to structured data, semi-structured and unstructured data are also included, e.g., special descriptions (in the form of audio, image and video files) from the community and social media
Veracity	The reliability (consistency, accuracy, trustworthiness, and availability) of data	Relatively low availability; MET data for all breeding lines and phenotypic data across target environments	High availability, e.g., omics data or historical climate and weather data across countries, regions and target environments
Variability	The extent, dimension, and speed at which the data are changing	From medium to high, regional or partially genome, e.g., METs with missing entries or data or variable phenotypic values and scores	From medium to high, global and whole genome, e.g., phenotyping traits at variable scales, omics data with variable depths and coverages, genotyping using different marker panels
Validity	Accuracy and correctness for the intended use	Case-dependent, changing with traits, marker types, and platforms	Case-dependent, changing with traits, data varieties, data generation platforms, and the evolution of those characteristics
Visibility	The visualization of data in a manner that is readable	Traditional graphs to plot data points	Clustering, tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees to represent data
Value	The actionable insight or functional knowledge extracted from data	Can be largely evaluated straightforwardly, and is dependent on individual case and data variety	Useful datasets can be extracted from big data, but value can usually only be determined through data mining and extensive processing
Vexing	The effectiveness of the model that can be developed from the data	Models can be developed largely using less-abundant phenotypic and genotypic data but with low effectiveness	Models can be structured and optimized using diverse data sources collected spatiotemporally at the levels of G-P-E, with relatively high effectiveness

Analyzing	Text analytics • Sentiment analysis • Face recognition • Voice analytics • Movement analytics • Clustering • Classification • Correlation and association • Model fitting and optimization
Artificial Intelligence	Robotics • Image analysis • Crop and soil monitoring • Machine learning • Modeling • Training software • Predictive analytics • Data mining • Data-to-knowledge • AI-based GS • iGEP
Impacts	Scale • Precision • Accuracy • Speed • Predictability • Efficiency • Genetic gain • Economic gain

Value

Figure 1. Big data properties associated with smart breeding and their impacts and value.

"Breeding events" represent the breeding activities from which big data are generated. "Datafication" refers to the processes by which subjects, objects, and practices are transformed into digital data. "Analyzing" refers to data analyses for breeding-related activities. "Artificial intelligence" (AI) represents the fields that need to be supported or assisted. Breeding-related big data have large impacts on breeding value through increased genetic and economic gain. Data properties are described by nine words beginning with V. G-P-E, genotypephenotype-envirotype; MET, multi-environmental trial; GS, genomic selection; iGEP, integrated genomic-enviromic prediction.

Properties	Conventional breeding	Smart breeding
Measurement	simple	complex
	data varieties described by simple words	data varieties described by both simple words and complex language
	largely empirical and less modeled	empirical, highly dependent on modeling and simulations
	relatively rough	relatively precise and accurate
	largely at the macro scale	at both the macro and micro scales
	less organized or standardized	well organized and standardized
Source	few or several	multiple
	largely focused on aboveground plant tissues	includes both above- and belowground plant tissues
	target environments	both target and nontarget environments
	largely normal environments	both normal and stressed environments
	plants	plants and their surroundings/environments
	plants as hosts	hosts and their companion organisms
	focused on target species	target species and their relatives
	pictures/photos as major images	pictures/photos, audio, video, and many more varieties of data
Collection	dispersed actions/events	dispersed and systematic actions and events
	manual + automatic	manual, automatic, and simulated
	observation by the naked eye with relatively simple facility support	observation by the naked eye with sophisticated facility support (e.g., remote sensors and robots)
Processing	less cleaning required	significant cleaning required owing to missing data and complicated structure
	easy integration	difficult integration
	relatively little transformation needed	various transformations required for data from different sources
	no need for regular reduction	dimensionality reduction often needed
	little discretization required	substantial discretization required
	all data utilized	subsampling required
Storage	notebooks, paper, spreadsheets, and databases	spreadsheets + databases, warehouses, the cloud, and local networks
	cumulative over the short term	cumulative and retrievable over the long term
	individual computers and networks	individual computers, sharable and connected networks, and the cloud
Sharing	occurs through notebooks, emails, and databases	occurs through warehouses, databases, networks, and the cloud
	low level	high level
	largely private	largely public and open
	largely passive	largely active
Analysis	largely manual	largely automatic and high throughput
	driven by human brainpower and computers	driven by human brainpower, computers, big data, and artificial intelligence
	largely delayed	largely in real time
	largely static	dynamic and visualized spatiotemporally
Mining	gene and genome structure	gene and genome structure and function
	individual cloning and functional analysis, like individual fishing	group cloning and functional analyses, like net fishing
	genetic metabolism and pathways	genetic pathways and networks
	simple gene modification and regulation	complex gene modification and regulation
	simple approaches	approaches supported by big data and artificial intelligence

Table 1. Comparison of data properties between conventional and smart breeding.

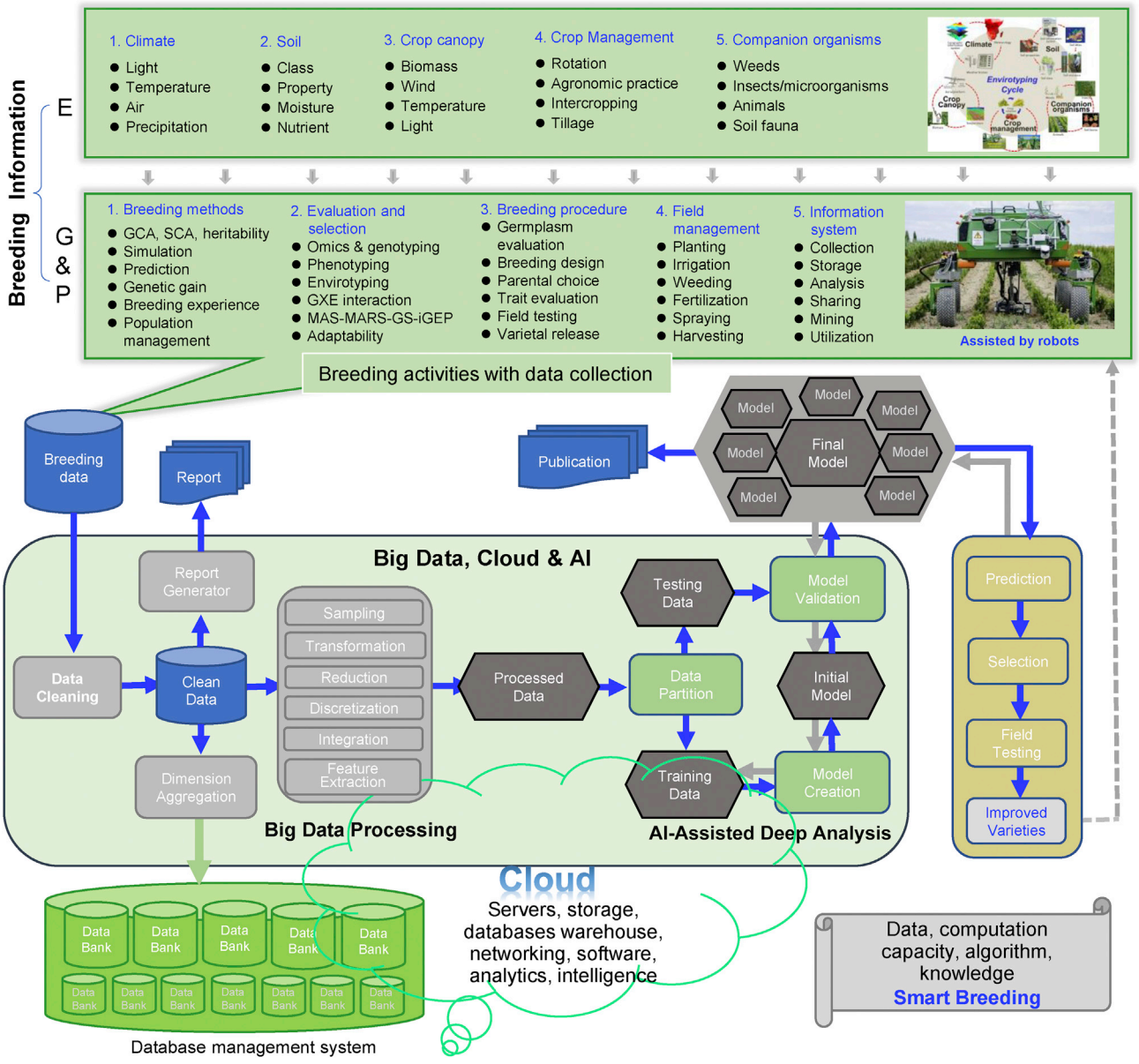

Figure 2. Overview of a system using big data and AI for smart breeding.

Data are collected for envirotype (E), genotype (G), and phenotype (P). Big data are stored, processed, and sampled to build and validate models with machine and deep learning and AI-assisted deep analyses. Trained models are used for phenotype prediction and selection to develop improved varieties. Smart breeding is driven by data, computational capacity, algorithms, and knowledge. GCA, general combining ability; SCA, specific combining ability; MAS, marker-assisted selection; MARS, marker-assisted recurrent selection; GS, genomic selection; iGEP, integrated genomic-enviromic prediction.

plant heights (Duvick, 2005; Denison, 2015). In general, the currently available HTP platforms are not comparable with visual observation, which enables close access to any individual in a high-density population. International phenotyping networks like IPPN and European plant phenotyping networks (Morisse et al., 2022) could help to coordinate phenotyping worldwide. Future robots, equipped with G-P-E big data and AI, could enable automatic, precision HTP for all target traits, rivaling human observation and measurement.

Genotyping

As the second "-typing" technology, genotyping has been extended to include sequencing, genotyping by sequencing, molecular profiling, and various other approaches that identify genotypes. Over the past 40 years, genotyping technology has experienced significant evolution in methods from systems based on gels (G1) to those based on fluorescence (G2), solid chips (G3), and liquid chips (G4), and finally to whole-genome sequencing by ultra-automatic sequencing facilities (G5) (Xu et al., 2020b). Technical revolution has significantly improved the throughput (from single to millions of markers at a time), resolution (from a 10–30-cM interval per marker to many markers per gene), and cost (from several dollars to less than one cent per data point). Genotypic information has been expanded from DNA to RNA (identifying genes and gene expression

levels; Boyle et al., 2008), proteins, and networks (linking genes to specific pathways), and can thus include all data that distinguish between genotypes, such as full sequences and their variations, multiomic variants, gene expression levels, and haplotypes.

Three representative genotyping platforms that evolved over the past four decades are gel-based RFLPs, fluorescence-based SSRs, and chip-based SNPs for application of different marker types; KASP and TaqMan platforms for single-marker genotyping; and Illumina and Thermo Fisher systems for chip-based high-density genotyping. In addition, genotyping by sequencing has been used to randomly capture a large number of markers across the whole genome (e.g., Elshire et al., 2011). As a more recent genotyping technology, genotyping by target sequencing combined with capture-in-solution (liquid chip) is a high-throughput, automatic, and large-scale platform developed to capture multiple SNPs (mSNPs) from single amplicons (Guo et al., 2021). From a set of 40K maize mSNPs, multiple panels with various marker densities (1K–40K mSNPs) could be developed by sequencing at different depths. Two alternative approaches, multiplexing PCR and probe hybridization, can be used to generate a wide range of marker scales required for different predictive breeding programs. Genotyping by target sequencing has fixed target SNPs so that genotypic information can be easily handled, accumulated, compared, and shared.

Envirotyping

The third "-typing" technology, envirotyping, refers to collecting and capturing all environmental factors that could affect plants and their phenotypes. The concept was first proposed at international conferences as e-typing or environmental assay in 2011 and 2012, followed by formal publications (Xu et al., 2012; Xu, 2015) and full discussion as envirotyping (Xu, 2016). The term "envirotyping" has also been used to refer to the collective body of methodologies for characterizing environments within METs and frequent repeatable environment types (Cooper et al., 2014b, 2016). Envirotyping is considered one of the seven most popular trends in plant improvement since the 1990s (Bernardo, 2016). A recent literature search for the term "envirotype" identified two references that were unintentionally missed in our previous publications. The term was first introduced in ecosystems research by Patten (1998), who considered that the direct genotype–phenotype model of classical genetics was incomplete and an external envirotype was needed to complete the mechanism. Later, Beckers et al. (2009) adopted the term in the context of disease research in mice, arguing that incorporating envirotypes into experimental designs would be essential for the accurate modeling of human diseases. To match with genotyping and phenotyping, the term envirotyping is recommended to describe the measurement of envirotypic (enviromic) variables, including all micro- and macro-environmental factors that influence the growth and development of plants.

Plants are affected by both internal and external environments. Internal environments can be further divided into intercellular and intracellular, including temperature, moisture, nutrients, and chemicals. External environments are those that surround the plant and have direct effects, such as light, air, temperature, water, and soil. In a broad sense, crop management (e.g., crop rotation) and companion organisms around the plant (such as insects and weeds) can also be considered part of such environments (Xu, 2010) (Figure 2).

Soil, as a substrate for crop growth, is arguably one of the most complex ecosystems on Earth (Daniel, 2005) and is a significant component of the envirotype. Repeated tilling and cropping result in perpetually disturbed soil with highly altered microbial profiles. These changes are compounded by the effects of fertilization, irrigation, crop rotation, and other agronomic practices. Such disturbances cause plants to become maladapted to the ecology of modern agricultural soil. Researchers have sought to understand the genetics of microbial associations and how plant genotypes affect activities such as recruiting beneficial rhizobia and mycorrhizae, establishing microbial communities, excluding pathogens, and potentially influencing food quality (as reviewed by Wallace et al., 2018). Microbial profiling and metagenomic analyses of environmental and crop-associated microbial communities could be harnessed as a part of the envirotype for future predictive breeding.

Environmental variables can be classified based on their predictability, repeatability, and manageability. Mega-environmental factors determined by crop seasons and geographic location (longitude, latitude, and altitude), which include temperature, light, precipitation, and soil properties, are largely predictable across years, whereas others are less predictable. The environmental effects, GEIs, and EE interactions caused by major environmental variables are largely repeatable across experiments and years, but those caused by minor environmental variables are less repeatable. Some environmental variables, such as those associated with agronomic management practice or with managed or controlled environments, are manageable.

The concept of heritability for genotypes can be extended to envirotypes. The genotype has 100% heritability because it remains unchanged across environments and experiments, whereas the heritability for envirotypes can be defined as the proportion of environmental variance that can be repeatedly explained spatio-temporally. It is similar to "repeatability" and largely "determined" by mega- and major environmental variables that are predictable, repeatable, or manageable. Mega-environmental variables have almost zero variance and can be treated as fixed effects, whereas major or minor environmental variables may have non-zero variance and can be treated as fix or random effects, depending on their heritabilities.

If repeatable GEI patterns are identified, then the target region must be divided into subregions or mega-environments. Breeding and utilization of mega-environment-specific cultivars will convert the repeatable GEI into G within mega-environments, thereby improving heritability (selection reliability) and genetic gain. If no repeatable GEI is found, then the target region must be treated as a single mega-environment, and the GEI must be accommodated by adequate testing (Yan et al., 2022).

Envirotyping has received increasing attention with the establishment of advanced phenomics platforms and facilities, becoming a routine practice in controlled or managed environments. For example, the Netherlands Plant Eco-phenotyping Centre consists of six complementary, experimental modules, two of which were designed for HTP under simulated multi-environment climates. Shennong National Phenomics Research Facility (Wuhan, China) was designed to precisely monitor and control light, temperature, water, soil, insects, and diseases.

There are additional reasons why envirotyping and its applications should receive more attention, including the six aspects described previously (Xu, 2016). Here we discuss only those that affect iGEP. First, in addition to the overall performance of specific varieties across environments, breeders care about not only the ranks of varieties within the environments but also the environmental effects. Major environmental factors can be identified and used to classify environments, which generally also contribute to GEI. Second, improved envirotyping can help better understand genes, pathways, and networks; crop performance under abiotic and biotic stresses can be better predicted using predictable mega-environmental factors. Third, by controlling environmental variation, environmental profiles can be built, relevant factors can be dissected into functional components, and EIs can be constructed. Fourth, by characterizing and classifying environments or experimental sites, we can optimize MET sites and determine target environments or market segments for the release of improved crop varieties. Fifth, envirotyping and enviromics enable better understanding of GEI. GEI components can be dissected, and the effects of crop management on gene expression (agronomic genomics) can be determined. All these five aspects will contribute to improving prediction accuracy and optimizing models and parameters in iGEP.

Multidimensional breeding information

In early plant breeding history, breeding data included only one dimension (phenotype). Although METs and field tests have collected environmental data in MSEs for many years, this third dimension (envirotype) has only been used more recently (Costa-Neto et al., 2021a; Cooper and Messina, 2021). In the era of smart breeding, a significant proportion of big data will be collected by tri-typing technologies across time and space (developmental stages, METs, and years). These multidimensional data, in addition to G-P-E, have become increasingly important for breeding across various environments and ecological conditions.

Challenges in breeding with big data

MSEs have been partitioned into functional components to enable optimal execution; this starts from new breeding tools/technologies, moves forward to line development and commercial or product development, and ends with value chain development (Eathington et al., 2007). There are significant differences between conventional and smart breeding pipelines (Table 2). A typical MSE usually runs at the following scales each year: 2K–5K populations, 200K–500K segregating lines, 10K–30K finished lines, 200K–500K test crosses, 20K–50K finished hybrids, 1–2M nursery rows, and 2–4M yield test plots (David Butruille, 2006, personal communication). Tomato breeders are evaluating approximately 35K genes individually and more than 610M cases of gene-to-gene interaction systemically to develop superior cultivars (Lee, 2020). Continued success in breeding depends on developing novel algorithmic and computing system approaches and evaluating these methods in agricultural settings. For storing, transferring, analyzing, and sharing data internally, an efficient data management system must be built and integrated with other internal services.

The big data analytics process comprises three major components—data capture, data analysis, and interpretation—that lead to insights into the underlying biology, optimized genetic models, and breeding decisions. There is a risk of drowning in the massive

Pipeline	Conventional breeding	Smart breeding
Breeding design	genotype, phenotype, and pedigree information can be used to select parental lines and design crosses and breeding schemes based on breeding objectives	all genotype-phenotype-envirotype big data and established models can be used to identify the best crosses and the most appropriate breeding scheme; crop varieties can be designed at the micro and macro scales using all available information
Breeding procedure	organized, designed, and standardized breeding procedures can be developed based on breeding practices and the limited information available	highly organized, designed, and standardized breeding procedures can be developed with assistance from big data and artificial intelligence through multi-disciplinary collaborations
Selection	based on favorable recombinants; largely focused on major genes; highly efficient for target environments; assisted by direct selection indices	based on design; focused on both major and minor genes; highly efficient for both target and nontarget environments; assisted by both direct and indirect selection indices (including modeling and prediction)
Multi-environmental trials	site selection with less optimization; accumulated data and observation are given relatively high weight; managed with normal production practices	site selection optimized based on accumulated multi-environmental trials and environmental data with future perspectives; well managed using precision agriculture with environmental data collected and accumulated
Varietal release and commercialization	additional field testing and trials are needed to determine target regions or market segments after multi-environmental trials	targeted market segments for varietal commercialization are determined efficiently and precisely using big data, artificial intelligence, and model prediction
Plant variety protection	molecular marker numbers used as a major criterion, with a small number of markers; both phenotype and molecular profiles used, with pedigree information involved	genetic similarity indices constructed using high-density marker and omics data; molecular profiles used as criteria along with phenotypes; pedigree and parental contribution determined and derived based on marker and omics data or profiles

Table 2. Comparison of conventional and smart breeding pipelines.

amounts of data generated by automated tri-typing systems. The major challenge lies in data management and synergism across data collection and analyses (Coppens et al., 2017). Breeding data have diverse data formats, including raster, vector, topography, model, table, text document, statistical, map, and figure formats, largely characterized by their site- and season-specific collection, source diversity, and generation periodicity. They are usually difficult to process and analyze using regular statistics and methodologies (Table 1).

There are five key issues to be considered in future tri-typing and data collection. First, ontology and protocol systems: all data collection should follow the same ontology and protocols to maximize the capacity for data integration, archiving, and comparison. Such systems should be implemented with consistent funding to support the breeders' focus on the breeding target, as shown in maize for long-term selection of protein and oil contents (Moose et al., 2004) and in data collection over years and METs (Cooper and Messina, 2021). Second, facilities: they must be able to extend breeders' hands and senses to reach more complicated situations, provide standardized and consistent measurements, and generate uniform data. Third, throughput: methods for high-throughput data acquisition are necessary to collect data from a huge number of individuals within a specific time window automatically and efficiently. Fourth, precision and accuracy: "garbage in, garbage out," data precision, and accuracy are crucial for big data. Garbage data are not better than nothing and cause trouble in data mining and utilization. Fifth, individuals versus populations: data should be acquired in environments comparable with the field condition. Plant density significantly affects both phenotyping and envirotyping, owing to GEI. Therefore, phenotyping protocols established for single plants should be updated for populations grown at high densities. How these five key issues are addressed will determine the extent to which the collected phenotypic data can be used for predictive breeding.

Intelligent integration of G-P-E data

"Point-to-point" integrations were the norm until middleware, data integration platforms, and application programming interfaces became fashionable. To facilitate more efficient use of data, new technologies have emerged over time, including data warehouses (ETL [extract-transform-load]), data mapping (data relationships), semantic mapping (ontologies), data modeling (integrated databases), and data lakes (raw data storage) (Morgan, 2018).

Data can be integrated in different ways (Lund, 2020): uniform data access, common data storage (the cloud), application-based solutions (specialized programs to locate, retrieve, and integrate data), common user interfaces, and middleware data integration. Data integration can be categorized into three types. (1) Link integration begins the query with one data source and then follows links to related information in other sources. Software solutions to integrate multiomics data leverage data-mining algorithms to collate heterogeneous biological information from across the web into knowledge networks (Marsh et al., 2021). (2) View integration leaves the information in its source database but builds an environment around the databases. (3) A data warehouse or cloud brings all the data together under one "roof." Cloud-based solutions are very cost-effective, allowing fast, automated, and secure data integration (Lund, 2020). Multiple challenges arise when multiomics datasets are integrated intelligently. Some are more general to ML analysis, such as the presence of missing values or class imbalance, and existing reviews have already covered those subjects (Mirza et al., 2019; Song et al., 2020). Some are more specific, including the noisiness and complexity of multiomics datasets.

With each pre-processed dataset, the multiple datasets could simply be assembled with sample-wise concatenation and used as input for ML (known as "early integration"). There are three vertical integration approaches (Ritchie et al., 2015): concatenation-based integration combines datasets before analysis, transformation-based integration performs mapping or data transformation of each dataset before analyzing the transformed datasets, and model-based integration performs analysis separately on each dataset before combining the results (Picard et al., 2021). As different types of genomic data often have very large differences in scale, the data need to be transformed to a proper scale before integration (Table 1; Figure 1). On the other hand, prediction models can be established for each data type by multi-stage analysis, and individual models or results can then be integrated (Xu et al., 2020a; Xu, 2020). Deep learning (DL) algorithms can capture nonlinear patterns and integrate data from different sources (Montesinos-López et al., 2021a).

Convolution and pooling are two data integration strategies for combining small datasets into big ones. As almost all ML models need to work with very large datasets, the first strategy is to combine relatively small datasets to form a large dataset (López et al., 2022). As a mathematical operation, convolution merges two functions (sets of information) to produce a third one as a modified (filtered) version of one of the original functions. A pooling operation, such as downsampling or subsampling in ML, is used in convolutional neural networks (CNNs) to bypass the need to explicitly define which independent variables (inputs) should be included or selected for the analysis by optimizing a complete end-to-end process to map data samples (Wang et al., 2019c). Maximum pooling and average pooling are the two most popular pooling operations. The former performs DR and denoising, whereas the latter mostly performs DR (López et al., 2022).

Data integration among systems has been traditionally difficult and expensive owing to the complexity of data formats, data types, and even the ways in which data are organized. Multiomics datasets are noisy, sparse, and irregularly collected under diverse conditions and time points, resulting in heterogeneity, high dimensionality, and, thus, an ill-defined prediction. For example, multiomics data are hampered by the problem of a small number of observations and a large number of independent variables ("large p small n") (López et al., 2022). Therefore, integration can make the datasets large and comprehensive enough for iGEP. Making multiomics data findable, available, identifiable, and reusable is crucial for data reuse and discovery through good data management (Wilkinson et al., 2016), including integration. Development of international standards, ontologies, and vocabularies, including the breeding application programming interface, will ensure integration and interoperability across multiomics datasets (Selby et al., 2019). Knowledge networks and information hubs are used to centralize multiomics data, and research should be extended to incorporate the growing knowledge base of GEI and

environmental data (Xu, 2016; Morais et al., 2019; Marsh et al., 2021). Cloud and web computing are essential for integrating data and analysis together (Marx, 2013).

Before data integration, the original data should be optimized by removing or adding variables (Kuhn and Johnson, 2013). An independent variable should be removed if it has zero or near-zero variance, as such a variable has a single unique value or very few values. For highly correlated or nearly perfectly correlated variables, only one should remain, as they all measure the same information, and such collinearity could inflate the parameter estimates. Models with less correlated variables can minimize unstable parameter estimates, numerical errors, and degraded prediction performance. Removing variables also reduces the computational resources required and may result in a more parsimonious and interpretable model. On the other hand, variables can be added by creating dummy variables from nominal or categorical inputs and a categorical variable through original data transformation (López et al., 2022).

For optimal information integration, it is recommended to include a substantial overlap of common entries for the best prediction. As an example, GP accuracies were evaluated in two independent prediction sets in combination with calibration sets. Including data from all selection cycles in model training yielded the best results because interactions between calibration and prediction sets, as well as the effects of different testers and specific years, were attenuated (Auinger et al., 2021).

Data integration platforms have focused on low-code and no-code tools that do not require specialized knowledge of query and programming languages, data management, data structure, or data integration (Morgan, 2018). Available data integration tools have been reviewed by Morgan (2018), and those most relevant to breeding data include ETL platforms to extract data from a data source, transform it into a common format, and load it to a target destination; data cleansing tools to identify, correct, or remove incomplete, incorrect, inaccurate, or irrelevant parts of the data; data warehouses to centralize data repositories; metadata management tools to enable the establishment of policies and processes that ensure information can be managed across the organization; data connectors to import or export data or convert them to another format; and data profiling tools to understand the data and their potential uses.

AI and robots in plant breeding

Modern breeding for mechanized, engineered, and facility agriculture introduces immense challenges. These challenges arise from increasing data volume, diverse samples, low valid-event rates, network complexity, heterogeneous data, and the need for data sharing (Figure 1), thus calling for AI. AI algorithms are used to create expert systems for prediction or classification based on input data. AI will have extensive impacts on breeding information systems. Breeders' experience and knowledge can be transferred into future AI-assisted breeding systems, and digitalization of breeding experience will promote the transition of breeding from empirically driven to AI-driven.

The combination of big data and AI has been referred to as both the fourth paradigm of science (Hey et al., 2009) and the fourth industrial revolution (Gil et al., 2014; Schwab, 2017). Buried deep within big data are immense opportunities for future plant breeding through AI (Figure 2). Plant breeding will be driven by AI in four ways: (1) AI-assisted breeding systems will play a significant role in theoretical study, evaluation, selection, breeding procedure development, and field management; (2) AI-equipped robots will interact with all the processes involved in data collection, storage, analysis, sharing, and utilization, significantly upgrading breeding information systems; (3) AI systems will benefit from historical experience and relevant knowledge produced and accumulated in breeding programs; and (4) breeding systems driven by big data and AI will have a great capacity for design and prediction through model simulation and optimization.

Robots equipped with AI have found success in solving scientific problems. For example, AlphaFold 2, an AI-robot algorithm, significantly outperformed other teams in a protein-folding contest at the Critical Assessment of Structure Prediction 14 in 2020, and it recently predicted the 3D structures of almost all proteins. It follows that AI technologies automated via robotics could facilitate breeding and crop production in many ways, including information capture (observation and identification of breeding-related data), data analysis (integration of data and empirical breeding information to build selection models), and breeding decisions (genotype selection for starting the next cycle of breeding) (Table 2). Applying a careful experimental design, using adequate biological replicates, and capturing as much information as possible about environmental heterogeneity within and across field sites are important steps in generating datasets with which to properly train a model for predictive breeding.

iGEP: CONCEPT

Genomic prediction and its current limitations

Recently, Bernardo (2021) comprehensively reviewed predictive breeding. The word "prediction" was first used in this context by Doxtator and Johnson's "Predictive breeding," begun in the 1930s, which was initially focused on developing superior double-cross maize hybrids (Jenkins, 1934). The advent of recurrent selection in the 1940s led to predictions for the next cycle of selection. To predict the performance of single-cross hybrids, genomic BLUP (GBLUP) was developed in 1994 (Bernardo, 1994). With the development of molecular markers, rapid-cycle recurrent selection became possible in the 1990s, and multiple regression was used to predict plant performance, enabling early selection based on predicted genetic values (Edwards and Johnson, 1994; Hospital et al., 1997; Eathington et al., 2007). After a landmark article on genome-wide selection (Meuwissen et al., 2001), prediction methods shifted from multiple regression with fixed marker effects to ridge regression and Bayesian models with random marker effects (Habier et al., 2011). Since then, GP has been widely used in both animals and plants (Crossa et al., 2017; Voss-Fels et al., 2019; Xu et al., 2020a, 2021a; Fu et al., 2022). Using the concept of genomics-assisted breeding, Varshney et al. (2021) proposed genomic breeding with the incorporation of GP.

Although molecular marker-based GP has been widely used, the currently available system has six potential limitations: it is suitable for highly related germplasms, which have historically

Prediction method	Model	Components	Data size	AI
GP across environments	$p = \boldsymbol{g} + e + ge$	$\boldsymbol{g}' = (g_1, g_2, g_3, \ldots, g_n)$	$\boldsymbol{g}$: 100K to 100M$^+$	not needed
iGEP with envirotypic data	$p = \boldsymbol{g} + \boldsymbol{e} + \boldsymbol{ge}$	$\boldsymbol{g}' = (g_1, g_2, g_3, \ldots, g_n)$ $\boldsymbol{e}' = (e_1, e_2, e_3, \ldots, e_l)$	$\boldsymbol{g}$: 100K to 100M$^+$ $\boldsymbol{e}$: 10K to 10M$^+$	preferable
iGEP with multiomic and envirotypic data	$p = \boldsymbol{G} + \boldsymbol{e} + \boldsymbol{Ge}$	$\boldsymbol{G}' = (\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \ldots, \boldsymbol{g}_n)$ $\boldsymbol{e}' = (e_1, e_2, e_3, \ldots, e_l)$	$\boldsymbol{G}$: $2^+\boldsymbol{g}$s $\boldsymbol{g}_j$: 100K to 100M$^+$ $\boldsymbol{e}$: 10K to 10M$^+$	preferable
iGEP with spatiotemporal multiomic and envirotypic data	$p = \boldsymbol{G} + \boldsymbol{E} + \boldsymbol{GE}$	$\boldsymbol{G}' = (\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \ldots, \boldsymbol{g}_n)$ $\boldsymbol{E}' = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_l)$	$\boldsymbol{G}$: $2^+\boldsymbol{g}$s $\boldsymbol{g}_j$: 100K to 100M$^+$ $\boldsymbol{E}$: $2^+\boldsymbol{e}$s $\boldsymbol{e}_k$: 10K to 10M$^+$	required for best efficiency
iGEP for multiple traits	$\boldsymbol{P} = \boldsymbol{G} + \boldsymbol{E} + \boldsymbol{GE}$	$\boldsymbol{P}' = (\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \ldots, \boldsymbol{p}_m)$ $\boldsymbol{G}' = (\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \ldots, \boldsymbol{g}_n)$ $\boldsymbol{E}' = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_l)$	$\boldsymbol{P}$: $10^+\boldsymbol{p}$s $\boldsymbol{p}_i$: 2^+ $\boldsymbol{G}$: $10^+\boldsymbol{g}$s $\boldsymbol{g}_j$: 100K to 100M$^+$ $\boldsymbol{E}$: $2^+\boldsymbol{e}$s $\boldsymbol{e}_k$: 10K to 10M$^+$	required for full function

Table 3. Predictive models integrating big data and artificial intelligence for smart breeding.

AI, artificial intelligence; GP, genomic prediction; iGEP, integrated genomic-enviromic prediction; K, thousand; M, million; p, $\boldsymbol{p}$, and $\boldsymbol{P}$, single, vector, and matrix variables for phenotype; g, $\boldsymbol{g}$, and $\boldsymbol{G}$, single, vector, and matrix variables for genotype; e, $\boldsymbol{e}$, and $\boldsymbol{E}$, single, vector, and matrix variables for envirotype. The first four models are proposed for single phenotypes. The single phenotypic variable p becomes $\boldsymbol{p}$ when multiple traits are involved and $\boldsymbol{P}$ when multiple traits are collected spatiotemporally (across multiple environments).

been used in predictive breeding (Bernardo, 2021); it is affected by many external factors; it is environment specific, with limited incorporation of envirotypic information; traditional models typically perform a linear regression analysis with clear assumptions and are unable to capture complex G-P relationships; it uses only genotypic data largely generated by molecular markers; and, finally, it optimizes models using limited historical data. Models using different breeding programs have rarely performed well (Wallace et al., 2018), and accuracy can drop rapidly, even beyond half-sib family structures (Beyene et al., 2015). On the other hand, current GP approaches largely rely on the G-P association. Many other important data layers that explain trait variation, including multiomics information, particularly enviromics, should be fully incorporated.

As a new strategy for GP, ML has been receiving increased attention. The most popular ML methods, such as random forest (RF) and support vector machines (SVMs), are very easy to implement because few hyperparameters need to be tuned, although more user intervention is needed to preprocess inputs manually. As a part of ML, however, DL is more model robust; it performs automatic feature engineering by learning through multilevel transformations, and it captures complex patterns more powerfully (Montesinos-López et al., 2021c).

iGEP methods

Accurately predicting and selecting the best lines and hybrids for specific environments relies on the ability to model complex systems from a web of G-P-E data. Multiomics data are a prominent example of such high-dimensional, heterogeneous datasets, and they have complex multilevel structures, contributing to difficulties in model construction and optimization. GP models involve several complexities, as model selection depends on several factors, including trait genetic architecture, marker density, sample size, linkage disequilibrium size, and GEI. Among the factors that affect GP, models are the only one that determines the predictability when a set of training data and breeding populations are given.

Here, we propose a next-generation prediction strategy, iGEP. The phenomic data used to upgrade current GP models include agronomic traits as well as molecular phenotypes that are measured at the molecular level, e.g., gene expression with the transcript abundances of thousands of genes (Jansen and Nap 2001; Brem et al., 2002; Schadt et al., 2003), and the enviromic data involve cellular, intercellular, and external environmental factors. Variables used in predictive model construction can be upgraded from single dimensions (the vectors $\boldsymbol{p}$, $\boldsymbol{g}$, $\boldsymbol{e}$) to multiple dimensions (the matrices $\boldsymbol{P}$, $\boldsymbol{G}$, $\boldsymbol{E}$) (Table 3).

In the context of predictive breeding, iGEP can be classified into five categories based on the variables included in the model (Table 3). First is conventional GEI, i.e., GP across environments using marker data and the phenotypic data collected across environments. Conventional GEI is included to understand the contribution of GEI to the phenotype being predicted. Second are predictive models with envirotypic information incorporated. Here, vectors of enviromic data are included in the model to understand the contribution of each environmental factor and its interaction with genotype. Third is iGEP, which uses both multiomics and enviromics data. Matrices of genomic data and vectors of enviromic data are included in the model to understand the contribution of a specific environmental factor and its interaction with all omics factors. Fourth is iGEP using multiomics and enviromics data for all relevant environmental factors collected across time and space. This method uses matrices of multiomic and spatiotemporal enviromic data in the model to understand the contribution of all environmental factors and their interactions with all omics factors to phenotypes across time and space. Fifth is iGEP for simultaneous prediction of multiple traits using all G-P-E data collected across multiple environments.

Genomic prediction across environments

Several examples have demonstrated the potential for enhancing prediction of yield and other complex traits by including GEI effects in prediction models (Burgueño et al., 2012; Jarquín et al., 2014; Acosta Pech et al., 2017; Cooper and Messina, 2021). In wheat, modeling GEI using information on markers or pedigrees could enhance prediction accuracy (Burgueño et al., 2012). Using interactions between markers and environmental covariates to account for GEI, the prediction accuracy was substantially higher (17%–34%) than

that of models based on main effects only (Jarquín et al., 2014). In maize, using 2724 hybrids evaluated for three traits in 58 environments, genomic models that included the interaction of general and specific combining ability with environments increased the prediction accuracy by 12%–22% (Acosta Pech et al., 2017). More recently, a decentralized durum wheat trial distributed across the Ethiopian highlands showed that data-driven decentralized breeding could double the prediction accuracy and enhance local adaptation and superior yield performance (De Sousa et al., 2021). As phenotypic data collected across environments become increasingly available, incorporating GEIs into prediction should become a routine practice.

iGEP with envirotypic data

Identifying and harnessing suitable sets of coordinated genotypic and envirotypic predictors will provide new opportunities for predicting the consequences of GEI. Several approaches have been proposed for incorporating environmental data into GS to improve prediction (Heslot et al., 2014; Jarquín et al., 2014; Cooper et al., 2016; Li et al., 2018; Millet et al., 2019; Costa-Neto et al., 2021a, b; Li et al., 2021b). However, the environmental dimension has been explicitly addressed in very few instances. In an early attempt, day length and temperature were used as examples to generate an EI to enable integrated modeling and prediction (Li et al., 2018; Guo et al., 2020). Three major crops (maize, wheat, and oat) were used, with 35 212 phenotypic values for flowering time, plant height, and grain yield across 51 year–location combinations. Identifying such an EI enabled the GS of complex traits with an explicit environmental dimension (Li et al., 2021b). With a quantitative EI established, the observed phenotype with this EI can be modeled, and phenotypic performance in new environments can be predicted using historical weather averages, in-season weather, or forecasted weather data. Along with this general framework, an analytical package, CERIS-JGRA, was developed to enhance GP. However, more complicated models should be developed to adapt prediction models to complex EIs or a complete envirotypic profile including all environmental variables.

Two MET maize datasets were used to investigate new kernel models involving genomic and nongenomic sources of variation (Costa-Neto et al., 2021a). A total of 16 environmental factors were used to create an envirotype covariable matrix, which was added to five whole-GP models involving environmental covariables and their interactions. The models were tested under three prediction scenarios: newly developed hybrids, sparse MET conditions, and new environments. Gaussian kernels and deep kernels are more efficient at translating model complexity into accuracy and are more suitable for including dominance and reaction-norm effects (Costa-Neto et al., 2021a). As a toolkit, EnvRtype was developed to offer remote sensing tools for collecting and processing ecophysiological variables from raw environmental data; environmental characterization by envirotyping and profiling environmental quality; and identification of enviromic similarity for an enviromic-based kernel. Envirotyping parameters were fine-tuned for each plant species and target environment by literature mining (Costa-Neto et al., 2021b). EnvRtype, CERIS-JGRA, and other similar software packages represent cost-effective envirotyping pipelines capable of utilizing high-quality enviromic data for a diverse GP set.

iGEP can be improved by incorporating diverse but well-managed environments. Diverse environments consist of both predictable and unpredictable environmental variables, whereas well-managed environments can form near-isogenic lines (NIEs) that significantly differ in only one major, predictable variable (Xu, 2016). Comparison of diverse environments and NIEs provides opportunities for detailed dissection of phenotypic differences caused by complex environmental variation. A single set of NIEs can be used to characterize the contribution of each major environmental variable to specific phenotypic variations. Such analysis can be used to improve model construction in iGEP. The same applies for the following discussion involving multiomics data.

iGEP with multiomics and enviromics data

Multiomics data enable the generation of a global profile of a plant, including genomic sequence (genomics), DNA methylation (epigenomics), gene expression (transcriptomics), protein abundance (proteomics), gene translation (translatomics), metabolic flux (metabolomics), genotypic and environmental contributions to phenotypic variation (phenomics and enviromics), and metadata (Harfouche et al., 2019; Wu et al., 2021). Some sources of omics data have much more complex data structures than marker data, and multiomics data, being multidimensional, already complicate model construction and fitness. High-level omics data, such as gene expression or metabolite concentrations, can capture additive and epistatic signals from multiple genetic loci owing to their molecular proximity to macro-scale phenotypes. For example, transcriptomic and metabolomic data can be treated as an intermediate phenotype, endophenotype, or something close to a genotype to best model all the data. Predictive models developed with data from multiomics layers are expected to provide better prediction than the use of molecular markers alone.

Integration of multiomics data into GS models improves prediction accuracy by efficiently capturing minor and nonadditive effects (Westhues et al., 2017; Schrag et al., 2018; Xu et al., 2020a). In two early reports, using a combination of parental genetic and metabolic markers significantly improved predictions for biomass heterosis in *Arabidopsis* (Gärtner et al., 2009) but not for general combining ability in maize (Riedelsheimer et al., 2012). Later studies using genomic, transcriptomic, and metabolic data from larger datasets improved predictions of complex traits in maize (Guo et al., 2016; Westhues et al., 2017; Schrag et al., 2018; Hu et al., 2019; Yang et al., 2022) and rice (Wang et al., 2019b; Xu et al., 2021b). To improve predictions, multilayered LASSO was developed to enable learning of three layers of genetic features (genome, transcriptome, and metabolome) (Hu et al., 2019). Incorporation of higher-order gene interactions significantly improved the predictability of rice yield from 0.159 (GP alone) to 0.245 (multilayered LASSO).

Another major category of genotypic data is molecular interaction networks. The whole set of molecular interactions that occur within a particular cell is defined as the "interactome." An integrated strategy proposed for combining the interactome with ML involves mining information hidden in big data to identify the genetic models or networks (Wu et al., 2021). The strategy includes seven steps, three of which are related to training and prediction: model training and construction using basic

algorithms (linear regression, logistic regression, naive Bayes, SVM, and decision tree), evaluation to find the best models, and novel gene prediction. As a new dimension of omics data, interaction networks can be incorporated into iGEP to improve predictive breeding.

iGEP models need to include higher-dimensional profiles to accommodate enviromics information collected across time and space. For example, mega GEIs can be revealed with temporal, spatial, and climatype data using models such as: Space × Time × Genome × Mega-genome × Metagenome. Prediction with such complicated variables will require more AI support and higher computational capacity.

iGEP for multiple traits

The single phenotypic variable (*p*) becomes a single-dimension vector (***p***) when multiple traits are considered and a multidimension matrix (***P***) when multiple traits are observed across multiple environments (Table 3). Correlation and similarity among multiple traits should be used to develop iGEP models. Recent progress in multi-trait GP enhances accuracy in multi-year wheat breeding trials (Montesinos-López et al., 2021b). A multi-trait GP based on a multivariate linear mixed-effect model could efficiently leverage thousands of traits at once (Runcie et al., 2021). Using maize as an example, a multi-trait predictive breeding strategy was developed (Yang et al., 2022). The similarity between genomic predicted and observed values could be trained using agronomic, transcriptomic, and metabolic traits and then used to predict multi-trait similarity among the target traits of predicted objects (inbreds or hybrids).

When multiple traits are included as response variables, the model becomes a multi-trait GBLUP, which can be developed using GBLUP and Bayesian methods (as reviewed by McGowan et al., 2022). A multi-trait GS model can be expanded to a multi-trait and multi-environment (MTME) Bayesian model or realized through nonlinear frameworks such as ML. Multi-trait GBLUP is expected to improve prediction accuracy by enabling information to be borrowed among correlated traits (Chen et al., 2023). There is evidence that the larger the correlation between traits, the better the prediction performance of multi-trait analysis (Jia and Jannink, 2012; Jiang et al., 2015). Incorporation of environmental variables into iGEP must take into account the correlations among multi-environment variables and the contribution of each factor to total environmental variance and the predictability of environmental variables.

Challenges in iGEP

Feature selection and DR

Big data in predictive breeding are characterized by high dimensionality, which refers to both the sample size and the number of variables and structures (Xu, 2020). High dimensionality poses challenges to computation and analysis, even with AI-assisted data analytics. As the number of features or dimensions increases, the amount of data required for accurate generalization increases exponentially, a phenomenon known as the curse of dimensionality (Bellman, 1961). Traditional algorithms can become unstable with a large number of variables, which also contributes to false positives owing to multiplicity of statistical testing. Because of the large number of genetic variants, it is not generally feasible to use the data matrices directly in standard statistical analyses. To avoid the curse of dimensionality, feature selection (FS) and DR methods are used (Figure 2).

FS yields a subset of features from the original set that best represent the data, whereas DR transforms the features into a lower dimension. There are a large number of FS methods available, such as clustering, linear transformations (principal component analysis [PCA], singular value decomposition), spectral transformations, and convolutions of kernels (Gabur et al., 2022). In general, unsupervised FS methods are less prone to overfitting (Guyon and Elissee, 2003) and have the ability to improve predictions. The combined prediction models and FS methods have been compared for microarray datasets (Bolón-Canedo et al., 2014; Bommert et al., 2020), text analysis (Forman, 2003), and image interpretation (Dy et al., 2003). In soybean, a recursive feature elimination approach was used to reduce the dimensionality of hyperspectral reflectance data, resulting in considerably decreased computation time and enhanced prediction accuracy, especially for nonlinear learners (Yoosefzadeh-Najafabadi et al., 2021).

High dimensionality greatly challenges traditional statistical theory (Fan and Li, 2006). Owing to the size and complexity involved, the associated mathematical theory can differ from the traditional approach. Methods such as PCA may therefore not be feasible, and more efficient DR methods such as random projection based on the Johnson–Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984) should be used. FS methods were included in ML, along with nonlinear DR and random selection using 80% of the input dataset, and prediction accuracies were evaluated with the remaining 20%. Features obtained from FS filter methods were combined with rrBLUP, LASSO regression, gradient boosting machine (GBM), ANN, and RF predictors. Using a spring *Brassica napus* population composed of 950 F1 hybrids (for seed and oil yields) and a diversity collection of 191 wheat cultivars (for grain yield traits), ML methods outperformed current approaches, increased prediction accuracy, drastically decreased computation time, and improved detection of important alleles involved in qualitative or quantitative traits (Gabur et al., 2022). Therefore, FS methods and modern prediction models can be combined to effectively exploit cost-efficient genotyping data and improve prediction accuracies.

Aside from DR, various other approaches can be used to reduce the number of variables to make big data more manageable. For example, a large number of SNPs can be transferred into a significantly smaller number of marker bins (An et al., 2020). Alternatively, a subset of markers can be extracted by biological filtering (Kim et al., 2020). For a large number of environments, environment-related variables can be compressed into one variable by coding environmental effects to identify complex interactions (Li et al., 2022b). It is also possible to pinpoint the major environmental factors that can be used in DR. For well-established gene networks, key genes that regulate a network can be identified to form a subset of multiomics data. In addition, unique plant materials, such as NILs, and well-managed environments, such as NIEs, can be used to develop strategies for reducing dimensions and variables. Dimensions can also be reduced by using correlated high-heritability traits or high-heritability secondary traits to replace

low-heritability traits and by using an index constructed for multiple traits.

Model overfitting and underfitting

As the complexity of the ML model increases, the bias between the expected prediction and the true observed values is reduced and the error variance (change in the estimate with different training datasets) increases (López et al., 2022). In this case, two major issues occur when training ML models. Overfitting occurs when the developed model is too unique to the training set and thus loses the ability to generalize. The chances of overfitting increase with higher dimensionality of the training data. In ML, overfitting can cause some serious problems, including false relationships caused by noise, a complicated ML model for the availability of data, and unpredictable or poorly predicted models (López et al., 2022). If the ML fails, some approaches can be used, including those proposed by Shalev-Shwartz and Ben-David (2014): (1) increase the training sample size; (2) modify the hypothesis and change the parameters used; (3) change and optimize the feature representation, e.g., by FS and DR; and (4) change the ML algorithm. ANN is a nonparametric ML model that is very flexible and is subject to overfitting of training data. This problem can be addressed by dropping out (setting to zero) the weights of a certain percentage of hidden units, while extreme gradient boosting (XGBoost) controls overfitting by using a more regularized model formalization (Chen and Guestrin, 2016).

Underfitting occurs when only a few predictors are included in the ML model, which then poorly represents the complete structure of the data pattern. This problem also arises when the training dataset is too small to represent the population (López et al., 2022). Increasing and accumulating training data are critical for prediction accuracy and can minimize overfitting and underfitting issues. A balance between overfitting and underfitting must be maintained during model design. If the training dataset is not large enough, the trained model is likely to accurately predict the targets in the training dataset but fail when used to generate predictions.

Modeling with data complexity

Modeling iGEP faces increasing challenges of data complexity. First, metabolites, transcripts, near-infrared wavelengths, or hyperspectral wavelengths can serve as agronomic traits or features in place of SNP markers in GBLUP to estimate relatedness (Bernardo, 2021), complicating data origins, structures, and formats. Second, enviromic data increasingly change with management practices, such as application of growth regulators, water, and fertilizers to the target environment. It is important that models are capable of employing enviromic data to predict which practices are optimal for breeding objectives. Third, not every sample is perfectly clean. Interaction network data, in particular, commonly include incomplete, missing, or unbalanced data. Such imperfections can be improved with the accumulation of complementary data that have similar patterns.

For iGEP, strategies need to be developed for modeling situations with complicated, big, and changing datasets. Alternatively, a subset of samples can be used to predict the remaining samples using the leverage effect (to generate a big return by predicting with a small subset of samples for training). For example, two sets of parental lines (with N1 and N2 parents, respectively) can be used to produce N1 × N2 hybrids. Some hybrids can then be selected to train a model to predict the rest. With the accumulation of more populations, the best training population can be developed from the large pool. A GS4.0 breeding program using the leverage effect was recently proposed to combine doubled haploid (DH) breeding with GS to fit various breeding schemes in maize (Fu et al., 2022).

Extending the use of omics and systems biology approaches is necessary to understand complex biological processes and their interactions with the environment (i.e., GEIs) through the integration of multiple datasets at the level of a defined system (Pazhamala et al., 2021). Notably, these components of data complexity are under the influence of environmental changes (Figure 2). Using a comprehensive map of rice QTNs, a genome navigation system, RiceNavi, has been developed for QTN pyramiding and breeding route optimization (Wei et al., 2021). However, because of the limited information available on potential genetic interactions, outcomes cannot be precisely predicted for pyramiding QTNs that underlie the same trait. Therefore, a better understanding of genetic networks for the target traits, such as QTN-QTN and QTN-E interactions, will improve the precision.

iGEP: IMPLEMENTATION

AI in iGEP

In traditional statistics, genotypic values can be predicted by linear mixed models (GBLUP, rrBLUP), penalized regressions (ridge, Lasso), and Bayesian methods (A, B, Cπ, R) (reviewed by De los Campos et al., 2013). AI-assisted GP involves ML techniques, including kernel methods, SVM, RF, and all kinds of modern machine and deep learners. The philosophy behind the techniques includes bias-variance trade-offs, training and test sets, cross validation, matrices, penalties, priors, kernels, trees, and deep learners (forward in López et al., 2022). Integrated AI models enable the analysis of an organism in a dynamic multiomic fashion. The development of more sophisticated AI algorithms, such as iterative RF (iRF), will enable the creation of new integrated discovery spaces (Harfouche et al., 2019). AI, coupled with high-resolution, high-throughput, and field-scale phenotyping and envirotyping technologies, will serve as a key tool for understanding GEI and unlocking greater potential for iGEP.

AI: Models

The predictive models summarized in Table 3 can be implemented by developing specific statistical models for each case and by integrating G-P-E data with AI. As AI-based predictions, ML classifiers fall into three primary categories. (1) Supervised learning uses labeled datasets (genotypes and phenotypes) to train algorithms. When the labeled datasets are fed into the model, it adjusts its weights until the model has been fitted appropriately. (2) Unsupervised learning analyzes and clusters unlabeled (for example, unphenotyped) datasets to discover hidden patterns. (3) As a happy medium between these approaches, semi-supervised learning uses a smaller labeled dataset to guide classification and FS from a larger, unlabeled dataset. In iGEP, some multiomics data may be suitable for modeling by supervised learning and others by un- or semi-supervised learning, by which, for example, environmental variables can be classified.

Envirotypic data are probably more suitable for semi-supervised learning models. Some envirotypic data can be collected for individual genotypes at the scale of experimental plots, whereas others can be collected only for the training and testing populations as a whole at the scale of the experimental station.

Many ML models are constructed to incorporate fixed effects, whereas others use random effects. As an extension of regression models, mixed-effects models allow the incorporation of random effects (López et al., 2022). Traditional prediction models consider environmental effects as a random effect, whereas other models consider them as fixed effects. Major or predictable environmental variables can be considered fixed effects, and unpredictable or minor environmental variables can be considered either fixed or random effects, depending on which kind of model is used.

ML models are commonly classified based on parameter properties (López et al., 2022). In parametric models, all the predictors take predetermined forms with the response. Linear, generalized linear, and nonlinear models are examples. In nonparametric models, the predictors are constructed according to information derived from the data rather than being predetermined with the response. Kernel regression and smoothing spline are two common ML models. In general, nonparametric ML models are very flexible and are better at fitting the data, particularly some phenotypic and enviromic data. In semiparametric models, one portion of the predictors does not take predetermined forms, whereas the other portion does.

In ML, a "kernel" usually refers to the kernel trick, a method of using a linear classifier to solve a nonlinear problem so that the original nonlinear observations can be mapped into a higher-dimensional space where they become separable (López et al., 2022). The kernel trick enables nonlinear versions of any linear algorithms to be built by replacing their independent variables (predictors) with a kernel function, giving them greater advantages, particularly for iGEP. Among the 11 advantages (López et al., 2022), 6 are most relevant to the development of efficient iGEP models: (1) interpretation as scalar products in high-dimensional spaces; (2) complexity controllable through regularization; (3) integration with classical methods for gene prioritization, prediction, and data integration; (4) ability to enable further improvements in the scalability of conventional ML methods and their versatility for working with heterogeneous variables; (5) flexibility and elegance for use in spatial analysis of classification problems; and (6) great promise for dealing with very large multidimensional datasets.

Different types of predictors can also be combined as in kernel methods, where different types of input variables serve to define multiple kernels, which translate to (transformed) GGI similarity matrices that structure genotypic random effects. For example, two random genotypic effects may be defined. The first one can be structured with an SNP-based relationship matrix and the second by a metabolite-based relationship matrix. Another example consists of adding a dominance kernel to a mixed model that already contains a kernel for the additive effects. To include epistatic effects of all orders in addition to additive effects, reproducing kernel Hilbert space (RKHP) models can be used.

In addition to mixed models and Bayesian methods, many ML models are powerful for capturing complex nonlinear patterns and have thus been explored for GP (Montesinos-López et al., 2022); for example, DL is used for training networks with multiple hidden layers, such as deep neural networks (DNNs). As "semi-parametric inference models," DL generalizes ANNs by stacking multiple processing hidden layers, each consisting of many neurons. The most popular topologies in DL are feedforward networks, recurrent neural networks, and CNNs (Montesinos-López et al., 2021a). DL approaches can be extended to any trait of interest for breeding with available high-resolution phenomics and enviromics imaging data.

Given a set of multidimensional data structures, no currently available AI methods can accurately identify combinatorial patterns within and across those structures. The most appropriate class of algorithm for fixing this problem is likely to be a DL approach, such as CNN, or a decision-tree-based approach, such as iRF (Ma et al., 2018; Harfouche et al., 2019; Abdollahi-Arpanahi et al., 2020; Yan et al., 2021b). The first step in the progression toward iGEP is to introduce AI algorithms that can not only be incorporated into predictive models but also used to expose the underlying rules that inform these predictions, offering meaningful insights for breeding.

To quantify how close the predicted values produced by an ML model are to the true values, a loss function is developed to measure the quality of the model output by computing a distance score between the observed and predicted values (Chollet and Allaire, 2017). A similar concept in ML is explainability, indicating that what happens in the model can be explained from input to output. Interpretable ML extracts associations through the correlations between features and outcomes (Azodi et al., 2020). Explainability for G2P prediction is relatively new, and new DL algorithms may have enhanced interpretability (Montesinos-López et al., 2021a).

AI: Prediction

Although it is not possible with most ML methods, DL models more efficiently incorporate large numbers of omics and G-P-E data in the same models, including the model types listed in Table 3, by: (1) naturally capturing nonadditive effects and complex relationships and interactions; (2) efficiently handling large data and raw data without any preprocessing; (3) allowing training models with many hidden layers and capturing very complex linear and nonlinear patterns involving many types of inputs, such as images; (4) designing specific topologies (DNNs) to deal with any type of data; and (5) significantly reducing the number of parameters by allowing parameter sharing and performing data compression (using the pooling operation) (Montesinos-López et al., 2021a).

ML methods more easily capture complex relationships for iGEP. Like nonlinear methods, ML architectures can also include multimodal data and data types that are not suited to simple tabular formats. For example, RF can capture patterns in high-dimensional data to deliver accurate predictions and can also take into account nonadditive effects (Heslot et al., 2012). It demonstrated superior performance compared with linear models such as Bayesian LASSO (Ornella et al., 2014) and rr-BLUP, depending on the genetic architecture of the predicted trait (Spindel et al., 2015). Other ML models that have shown

potential for prediction include CNNs and feed-forward DNNs that can outperform linear methods (Ma et al., 2018; Sandhu et al., 2020). Multi-trait DL models can help us understand the relationships between related traits for improved prediction (Montesinos-López et al., 2018), and ensemble models are powerful by combining multiple ML methods that may produce weaker predictions by themselves (Jubair and Domaratzki, 2019; Banerjee et al., 2020). DL is efficient for extracting representative features from large datasets and has the capacity to account for feature interaction effects in multidimensional G-P-E datasets.

Including multidimensional datasets increases the complexity of analysis exponentially, requiring algorithms that are able to uncover relationships between the data types and the target trait (Danilevicz et al., 2022). ML methods can help improve iGEP, as they enable computers to learn patterns that could be used for analysis, interpretation, prediction, and decision-making without being explicitly programmed. ML-based iGEP has the potential to transform methods for plant breeding owing to the four factors summarized by López et al. (2022): (1) the massive amounts of data being generated in breeding programs are now available for training ML models; (2) new technologies, such as sensors, satellite technology, and robotics, allow scientists to generate not only genomic data but also phenotypic and envirotypic data that can be used in the modeling process to increase performance; (3) increased computational power now allows complex ML models with larger datasets to be implemented in less time; and (4) user-friendly ML software has become available for implementing a great variety of ML models.

Multimodal DL models are composed of multiple models, each trained using a single input type (e.g., rainfall, soil measurements, genetic data, hyperspectral imagery), or a single model trained on concatenated multimodal data. The different modalities enrich the available features for model learning, contributing to an improved final prediction (Danilevicz et al., 2022). Use of DL in iGEP is attributed to its more powerful automatic feature extraction, greater data representation capability for dealing with high dimensionality, and ability to capture complex patterns (Montesinos-López et al., 2021c). It can be expected that DL methods will outperform conventional ML methods for iGEP, with progress in the following fields: increasing numbers of large datasets available; more computer resources available for tuning of large grids of hyperparameter combinations; spatiotemporal multiomics available as inputs and preprocessed using nonlinear models to retain complex patterns; exploration and implementation of transfer learning and reinforcement learning; exploration of deep generative models (generative adversarial networks and variational auto-encoder methods) to generate new inputs; increased expertise dedicated to model calibration; increasing data sharing and open-source breeding; and development of more user-friendly software.

AI: Applications

The incorporation of DL in breeding pipelines is in progress and has been used for predicting parental combinations in hybrid breeding programs (Khaki et al., 2020), modeling and predicting quantitative traits (Sadeghi-Tehran et al., 2019), genetic diversity and classification (Yang et al., 2019a), and GS (Montesinos-López et al., 2021b). However, prediction methods, including those based on ML, have difficulty carrying their experiences from one case to another. Transfer learning in this regard should focus on storing knowledge gained when training a particular prediction algorithm and then using this stored knowledge to solve another related problem (López et al., 2022). The transfer can be based on, for instance, feature, model, and relation (Niu et al., 2020). Transfer learning provides an effective solution for the large volume of high-quality labeled data and considerable computational resources required to train a large ML/DL model (Yan and Wang, 2022) and has thus been successfully applied to cross-species prediction and plant phenotyping (Moore et al., 2020; Nabwire et al., 2021). Large-scale omics datasets have been generated and annotated for only a very limited number of model plants, such as *Arabidopsis*, rice, and maize, and it is impractical to produce well-annotated training data for all nonmodel species (Yan and Wang, 2022). One possible solution is to use transfer learning to achieve cross-species prediction by considering conserved gene functions and pathways between evolutionarily related species (Cheng et al., 2021).

Nonlinear learners (GBM, ANN, and RF) outperformed linear learners (rrBLUP and LASSO) on average when relevant features were selected in prediction across training sets (Gabur et al., 2022). The prediction of hybrid grain yield in maize was better with a 20 hidden multilayer perceptron (MLP) model than with classical linear models (Khaki and Wang, 2019). In another study with datasets from yeast, wheat, and rice, six ML methods (elastic net, rrBLUP, LASSO regression, RF, GBM, and SVR) outperformed two classical statistical methods (Grinberg et al., 2019). In a comparison of 23 independent studies, nonlinear models outperformed linear ones in 47% of all studies that included GEI and in 56% of studies that ignored GEI (Montesinos-López et al., 2021a). In wheat, four ML methods, including SVR, kernel ridge regression, RF, and AdaBoost.R2, significantly outperformed GBLUP, single-step GBLUP, and BayesHE (Wang et al., 2022b); improved prediction accuracies were reported for neural networks (Pérez-Rodríguez et al., 2012; Ma et al., 2018) and kernel-based models (RKHS) (Pérez-Rodríguez et al., 2012); and RF and MLP were the best-performing ML models for prediction of spectral data.

In general, ML models outperformed four explored Bayesian models and required less computational time (Sandhu et al., 2021). In white wheat, multi-trait ML models performed better than GBLUP and Bayes B for cross-location predictions, but their advantages diminished when GEI was included (Sandhu et al., 2022a). For soybean yield, an RF algorithm had the highest performance among all individual algorithms tested, including MLP and SVM (Yoosefzadeh-Najafabadi et al., 2021).

As an ML technique for regression and classification, GBM uses the assembly of multiple weak learners to establish a strong model; thus, its prediction ability is significantly better than that of single models (Che et al., 2011). XGBoost, one of the implementations of GBM, outperforms DL in some tabular data problems (Zamani Joharestani et al., 2019). In three out of four wheat datasets, GBM outperformed a Bayesian GBLUP model (Montesinos-López et al., 2022). With a large dataset of inbred and hybrid maize lines, LightGBM exhibited superior performance in terms of prediction precision, model stability, and computing efficiency (Yan et al., 2021a). With data collected

from multiple years (2014–2017) across the US and Canada, including environmental predictors, two GBM methods (XGBoost and LightGBM) improved prediction accuracy for grain yield of new genotypes by up to 20% compared with models of linear random effects (Westhues et al., 2021). In soybean, XGBoost or RF outperformed DL models in 13 out of 14 sets of predictions (Gill et al., 2022). Results from both cattle and plants showed that GBM outperformed RF, ANN, and CNN (Ma et al., 2018; Azodi et al., 2019; Abdollahi-Arpanahi et al., 2020). In an extensive study with 6 linear and 6 nonlinear algorithms evaluated with data on 18 traits across 6 plant species, no overall winner was found (Azodi et al., 2019). However, the prediction using a combination of multiple ML methods did outperform that using conventional linear methods.

Using two training populations phenotyped in multiple years and genotyped with 40 368 markers, 17 GS models for complex traits were compared. Only a few significant differences were found between models, with SVMs reaching the highest accuracy of 0.56. The parametric models showed consistently moderate accuracy, with little advantage over nonparametric models within individual years, but the nonparametric models had slightly increased accuracy when combining years (Merrick and Carter, 2021). Using one simulated animal breeding dataset and three empirical maize breeding datasets from a commercial breeding program, several groups of supervised ML methods, including regularized regression and deep, ensemble, and instance-based learning algorithms, were compared (Lourenço et al., 2022). All the methods showed reasonably high predictive performance, but their relative predictive performance was both data and trait dependent, complicating and precluding omnibus comparative evaluations of the prediction methods and thus ruling out selection of one procedure for routine use.

Potential challenges for deploying ML for prediction include: (1) developing and implementing consistent protocols, (2) reducing dataset dimensionality, (3) reducing class representation imbalance, and (4) accounting for environmental variations between conditions for plant growth in the training and deployment datasets (Danilevicz et al., 2022). The use of multimodal models and other DL architectures, such as recurrent neural networks and graph neural networks, remains largely unexplored (Danilevicz et al., 2022). It is important to note that when the dataset is considerably large, it is better to randomly split it into three parts for training, validation (or tuning), and testing (López et al., 2022). With more and more accumulated G-P-E data, cross-validation and model optimization can be performed more efficiently for iGEP.

Spatiotemporal models in iGEP

When spatiotemporal multiomics data are incorporated into iGEP, the number of features (dimensions) grows, the amount of data required to generalize accurately grows exponentially, and it becomes difficult to generalize the model. Therefore, more training data are required. As enviromics data have been largely excluded from previous prediction models, our discussion will focus on how we should take envirotyping and environmental variables into account in our future predictions. Including environmental variables as the third key dimension in prediction, as shown in Table 3, significantly increases dimensionality, as these variables can be collected spatiotemporally. It complicates prediction models owing to many unpredictable contributors, such as climate change, and it introduces multidimensional interactions that must be addressed in the prediction.

Spatiotemporal omics

The omics data generated during the past three decades largely lack temporal and spatial information. There is an urgent need to link space and time using integrative and scalable G-P-E data to better understand plant phenotypes and associated metabolic processes (Munné-Bosch, 2022). Collecting spatiotemporal omics data has become increasingly important, and more multiomics data have begun to appear. For example, stereo-seq combines DNA nanoball-patterned arrays and tissue RNA capture to achieve large field-of-view spatial transcriptomics at a cellular resolution, enabling the dissection of spatial cell-type heterogeneity in mouse embryonic tissues (Chen et al., 2022a). As a specialized form of DNNs, CNNs can be used to analyze input data that contain some form of spatial structure (Goodfellow et al., 2016).

Spatial mining uses data related to experimental sites, locations, and geography to extract spatial relationships and measurements that are not explicitly stored in spatiotemporal databases. Temporal mining uses large quantities of implicit or explicit temporal data to extract information and temporal relationships, such as whether temporal data follow cyclic, random, or annual/seasonal variations, etc. In plants, it usually involves omics data collected across growth and developmental stages and across crop seasons and years.

Spatiotemporal models

Crop growth models and ML methods can be employed to integrate large vectors of primary environmental covariates into meaningful environmental characterizations (Resende et al., 2021; Diepenbrock et al., 2022). By scanning for the time window that gives the best prediction of mean environmental performance, useful temporal environmental variants can be identified (Li et al., 2022a). By way of analogy, these approaches have been called enviromic prediction or envirome-wide association studies (Piepho, 2022).

Spatiotemporal environmental variables can be dissected into the major variables collected from predictive mega- or macro-environments and minor ones collected from less- or unpredictive micro-environments. Including spatiotemporal envirotypic data as fixed effects can improve the fitness of the model and reduce the noise caused by treating predictive environmental factors as random effects, as the phenotype can be better predicted. As an example, for modeling spatiotemporal environmental variables, a half-diallel maize experiment with 35 families and 2367 hybrids was conducted at 17 locations in the US and 6 locations in managed-stress environments. Crop growth models linked to whole-genome prediction offered a predictive accuracy advantage compared with BayesA ($r = 0.43$ versus $r = 0.27$) (Diepenbrock et al., 2022).

iRF is designed to model across the multiple polytopic space and may provide one of the first tractable AI solutions for systems biology and an ML algorithm for iGEP. For example, iRF would be able to use spatiotemporal G-P-E data to predict the phenomic layer and identify sets of genes and environmental factors that affect each of the phenotypes and their combinations (Harfouche et al., 2019). In parallel, spatiotemporal microclimate data can be repeatedly collected with ground-based sensor

networks and used as a reference on which data analysis pipelines can be developed. AI-enabled iGEP algorithms can then be used to evaluate breeding decisions and predict which variety will show the best performance in field testing (Harfouche et al., 2019) or under specific environments characterized by envirotyping.

EI in predictive modeling

Using detailed environmental data arranged in a quantitative descriptor, such as a covariate matrix, analyses can be performed to dissect the GEI, model genotype-specific sensitivity to critical environmental factors, dissect QTL×E interaction components, incorporate environmental data to model the GEI reaction-norm, profile the environmental gradient in an experimental network, and generate environmental relationship matrices for prediction (Costa-Neto et al., 2021b). We have three options for handling GEI in iGEP: reducing or eliminating GEI, optimizing iGEP using GEI, and optimizing GEI and METs using environmental data. In general, environmental data are collected under specific environmental conditions. Environmental similarity is largely determined by mega and major environmental variables and will, in turn, determine the similarities and patterns of GEI and the accuracy of iGEP.

EI: Definition and general considerations

The environmental mean (t_j) was used by Li et al. (2021b) as an EI (regressor variable). Thus, the regression for the index can be written as $t_j = c_0 + c_1 x_j$, where x_j is the environmental variant for the jth environment (Piepho, 2022). One of the prediction methods for multiple traits or multiple environments is to construct a selection index as in quantitative genetics. An EI can be treated as a new trait for performance prediction and heritability estimation. The EI for a single environmental variant can be extended to comprise several environmental variants. The regression model was further extended to comprise several EIs, each predicted using a linear combination of several environmental variants (Piepho, 2022), as envirotypes can be dissected into components, as is done for genotypes, and then used to construct EIs.

When an envirotypic item is included explicitly to model iGEP, its mean is no longer zero. Although ML provides a catalog of different models and algorithms from which we try to find the one that best fits the G-P-E data, there is no universally best model, and a set of assumptions that works well in one domain may work poorly in another (Wolpert, 1996). Therefore, when envirotypic data are incorporated into iGEP, different models, algorithms, and sets of hyperparameters should be tested for each specific dataset so that the best model and prediction can be found. Multivariate models, such as PCA and modern approaches from AI, will likely allow better definition of enviromic variables for improved EI.

To identify EIs, the Critical Environmental Regressor through Informed Search algorithm was implemented by examining four environmental parameters (photoperiod, temperature, photothermal time, and photothermal ratio) and by genotyping maize, wheat, and oat. The average value of the environmental parameters was calculated for the window from the ith to the jth day after planting. The parameter–window combination with the strongest correlation was then chosen as the EI, enabling GS of complex traits with an explicit environmental dimension (Li et al., 2021b).

EI: Model-related methods

Mixed models have been generalized for MTME by defining a structure for the random G×T and GEI effects (forward in López et al., 2022). For phenotypic traits or environments, unstructured or factor analytic variance-covariance matrices can be chosen. When explicit environmental characterizations are available, relationships between environments can also be defined based on environmental similarities (Jarquín et al., 2014). Therefore, prediction models for EIs can be built by using the environmental similarities and various relationship matrices derived from all kinds of combinations of G-P-E and their interactions. Important environmental variables can be measured to provide suitable environmental predictors for envirotyping and to enhance prediction (Cooper et al., 2014a, b; Jarquín et al., 2014; Messina et al., 2018; Voss-Fels et al., 2019; Costa-Neto et al., 2021b; Resende et al., 2021). SVM can learn nonlinear decision surfaces and perform well in the presence of a large number of predictors, even with a small number of cases (López et al., 2022). This makes SVM very appealing for tackling environmental classification and EI construction. Considering the importance of high-throughput environmental data (Rogers et al., 2021), GEIs and environmental covariates, among other factors, should be incorporated into prediction models (Crossa et al., 2021). Through enviromic assembly, relatedness among field trials can be established and only the most representative set of experiments is used to train models, an approach called "enviromic + genomic prediction" (Crossa et al., 2021).

With improved proximal and remote sensor technologies, important environmental variables that determine G×E×M interactions have been quantified and measured, bringing a wide range of opportunities for accelerating crop improvement through enviromic technologies (Cooper et al., 2020; Peng et al., 2020; Kusmec et al., 2021) and enabling environment-specific prediction (Rogers et al., 2021). With genotypic and environmental predictors, an integrated view of G×E×M interactions can be predicted across all breeding program stages for selection and hybrid advancement (Cooper et al., 2014b). In this case, M, as a way to modify E, can simply be included in prediction models as an environmental variable (Harfouche et al., 2019).

Using simulated data, an index-based enviromics method (GIS–GEI) was developed. Because of its higher granular resolution, GIS–GEI allows for accurate identification of sites for their most appropriate genotypes, better definition of target environments with high genetic correlation to ensure selection gains across environments, and efficient determination of the best sites for future experiments (Resende et al., 2021). To probe the genetic underpinnings of climate adaptation for crop species, climatype data have been amassed for every square kilometer of land on Earth. Using the Summit supercomputer, each square kilometer was then compared with every other square kilometer to identify similar environments (Streich et al., 2020). These climatype data were combined with GPS coordinates associated with individual crop genotypes to project which genes and genetic interactions are associated with specific climatic conditions (Beans, 2020).

EI: Kernel approaches

Enviromic kernels can be constructed to capture the macro-environmental relatedness that shapes the phenotypic variation of relatives (Costa-Neto et al., 2020). To implement this modeling

approach, main functions were designed for constructing environmental relationship kernels using environmental information, integrating these kernels into statistical models that account for different structures capable of explaining the phenotypic variation, and fitting regression models that account for genomic and enviromic data (Costa-Neto et al., 2021b). This model can be extended to handle more complicated models involving environmental variables and iGEP. The task of classification is to classify different classes based on known input labels (supervised learning). One method is SVM, in which kernel methods are used. A user-specified kernel function (similarity function) adds another dimension to the dataset.

Factorial and spatiotemporal structure of plant breeding data

A factor structure is a correlational relationship among a number of variables. Plant breeding data involve many variables with a complex factorial and spatiotemporal structure of multiple layers. Therefore, plant predictive breeding must consider all these complications (Shahi et al., 2022), involving MTMEs of multiple layers (Figure 2).

Multiple traits and multiple environments

When low-heritability traits have at least moderate correlation with high-heritability traits, the predictability for the low-heritability traits could be strongly increased by using a multi-trait model (Jia and Jannink, 2012; Montesinos-López et al., 2016; Budhlakoti et al., 2019). The prediction accuracy for low-heritability key traits can be improved by using high-heritability secondary traits (Jia and Jannink, 2012; Muranty et al., 2015). Building on the strength of ML, an integrative multi-trait breeding strategy that uses target-oriented prioritization was performed, with up to 91% of the accuracy achieved for identifying a candidate that is phenotypically closest to an ideotype, or target variety (Yang et al., 2022). This strategy first learned the similarity between genomic-predicted values and measured phenotypic values and then predicted the degree of similarity between inbreds or hybrids and the target with respect to hundreds of traits.

When environmental information is available, a univariate GBLUP model can be extended as a Bayesian genomic MTME model by adding the interaction term (Montesinos-López et al., 2016). Large-scale correlations among traits evaluated across diverse environments can be used to train prediction models. The performance of 13 quality traits in wheat was predicted using 2 multi-trait models and 5 datasets based on field evaluations over 2 years (Ibba et al., 2020). The Bayesian MTME model helps capture the correlations among traits and among years, thus increasing prediction accuracy. A combinatorial optimization model was combined with an RF for predicting the yield performance of crossing testers and inbreds. When the model was designed to detect GEIs, the RF model was able to capture other types of linear and nonlinear effects (Ansarifar et al., 2020).

Multiple layers

To produce a model for important biological interactions, an algorithm should be developed to build an accurate prediction from multiomics data layers to identify the combinatoric interactive elements within and between those layers. In iGEP, the ''layer'' represents different categories of G-P-E. Taking envirotype as an example, the layer consists of the envirotypic data from weather, climate, canopy, agronomic practice, and accompanying organisms (Figure 2). With two hidden layers, neural networks can usually represent functions with any type of polytope shape. Empirical evidence shows that using more than two hidden layers can better capture nonlinear patterns and complex interactions (Chollet and Allaire, 2017).

To approximate any continuous function, MLPs are designed for solving problems that are not linearly separable; they contain input, output, and hidden layers, and the input layer receives the input signal for processing. The output layer performs the required task, such as prediction and classification. When only one hidden layer is present, the DL model becomes a conventional ANN model, but more than one hidden layer can better capture complex interactions, nonlinearities, and nonadditive effects.

Each layer performs nonlinear transformations, and connections among these layers are made using weights. Using too few or too many neurons in the hidden layers will result in underfitting or overfitting, respectively. For example, using 5000 markers as input variables to predict grain yield (a continuous outcome) means that we should have the same number of neurons but only 1 output layer. There is no unique and reliable rule for how to determine the required number of neurons in the hidden layers, as this number depends on input neuron number, training data (amount and quality), and learning task complexity (Lantz, 2015). To determine the required neuron number, two approaches can be taken. The backward approach begins with a very large number of neurons and evaluates their performance, then decreases the number of neurons until there is no more gain in reducing the testing error. The forward approach begins with half of the input neurons and increases the number of neurons until no significant gain is observed.

Hyperparameter tuning and dropout

Successful applications of ANN/DL depend on how the right hyperparameters are chosen to begin the learning process. As a critical aspect of the ML training process, hyperparameters need to be tuned for network topology, activation functions, hidden-layer number, neuron number in each layer, learning rate, etc. (López et al., 2022). Hyperparameter selection is performed with the goal that a model should neither underfit nor overfit the training datasets. However, this task is challenging because the number of hyperparameters required in ANN/DL is large. Tuning hyperparameters for DL models tends to be more computationally intensive (Danilevicz et al., 2022) and can be performed by grid search, random search, Latin hypercube sampling, and optimization (Koch et al., 2017). With more and more accumulated G-P-E data, hyperparameters need to be tuned more efficiently.

To prevent overfitting and improve the model's generalizability, regularization (penalization) is used to reduce testing errors so that the model performs well on new data as well as training data. One means of regularization is to minimize an augmented loss function. For ANN/DL models, more than one hyperparameter is needed, and different levels of penalties can be applied to different layers and hyperparameters. Another type of regularization is the dropout, which consists of setting to zero a random fraction (or percentage) of the weights of the input neurons or hidden neurons (Srivastava et al., 2014). The dropout (%) is determined based on starting point, network size, application layer, step size, and network weight (López et al., 2022). The

dropout method can be implemented with any type of loss function. All the loss functions can be converted to regularized (penalized) functions. It remains to be determined how loss functions and dropout can be used for enviromic data and the integration of different datasets.

SMART BREEDING BY PREDICTION-BASED CROP REDESIGN

Recent developments in genetics, genomics, and molecular biology have significantly accelerated the discovery of functional genes, metabolic pathways, and molecular networks. Multiomics information has quickly driven plant breeding from selection-based programs to prediction-based crop redesign, contributing to improved genetic gain through the creation and utilization of genetic variation. Smart breeding can be performed through iGEP to improve selection efficiency, accelerate the breeding process, breed new crops via *de novo* domestication, and design ideotypes through synthetic biology. In contrast to the iGEP methods discussed in previous sections, three strategies that may not require any proposed statistical models are pathway-driven breeding, *de novo* domestication, and synthetic biology (Wu et al., 2021). However, information from genes, pathways, and networks can be transformed into a new dimension of genotypic data and then incorporated into predictive models.

Breeding by crop redesign at the micro scale

Three models can be used to redesign crops at the micro scale (Table 4), including designs based on genes, metabolisms, and networks. As a model crop for gene design, gene cloning, and functional analysis, rice has provided deep insights into phytohormones and growth, nutrient use efficiency, responses to abiotic stresses, defense activation, and signaling in biotic interactions, photoperiodic flowering, and the control of fertility and sterility (Chen et al., 2022b). With functional analysis of increasing numbers of candidate genes, the best alleles, allele combinations, and favorable haplotypes can be identified, modeled, edited, designed, and used for marker-assisted selection (MAS) and prediction. Similarly, crop design was proposed to develop "smart super rice" (Qian, 2017) and "green super rice" (Yu et al., 2020; 2021b), identifying target genes controlling phenotypes of interest and germplasm resources where the relevant genes could be sourced.

At the micro scale, metabolic pathways can be substituted, modified, or improved through metabolism design. "The C4 Rice Project," an international collaboration to introduce C4 traits into rice, is expected to increase photosynthetic efficiency by 50% (https://c4rice.com). A pan-European research initiative, the CropBooster Program, aims to explore scientific options for improving plant performance by increasing photosynthesis (Harbinson et al., 2021). To address water limitations and mechanized rice production, elite rice varieties have been developed by introgressing drought tolerance and water-saving genes (and probably unspecified but associated pathways) from upland rice in an example of the network design strategy. The new rice can be planted by direct seeding in rainfed fields (Luo et al., 2019), triggering the rice "blue revolution" in China by freeing rice cultivation from irrigation, planting, and harvesting rice as wheat and reducing greenhouse gas emissions (Xia et al., 2022).

Multi-scale regulation of networks can be measured as changes in mRNA synthesis, stability, and decay and in protein translation, activity, affinity, and decay. The regulatory linkages across biological scales can be constitutive, tunable, or switchable under changing environments. Regulatory genomic variation of influential nodes in network modules perturbs network properties, such as hubs,

Levels	Models	Strategies	Examples
Micro scale	gene design	site-directed gene knockout, mutation, or gene editing; RNA interference; transgenics; marker-assisted selection	favorable alleles or haplotypes, allele/haplotype combinations generated and selected via marker-assisted selection
	metabolic design	substitution, modification, optimization, and improvement of metabolic pathways	metabolic pathways modified for improved photosynthetic rate
	network design	design and improvement of parameters, such as network regulators, network structure, network nodes, and borders	rainfed, direct-seeded, and drought-tolerant rice
Macro scale	individual design	morphology, ideotype, assimilate distribution, biotic and abiotic stress tolerance, trait interaction, and complementation	ideotype created by combining semi-dwarfism, erect top leaves, and strong stems
	population design	structural optimization, ecological stabilization, adaptability improvement, high-density planting, functional canopy, photosynthetic efficiency, and source-sink coordination and compensation	maize plants suitable for high-density planting and mechanized grain harvesting; small and miniaturized crops for facility agriculture and verticulture
	species design	integration of favorable traits from different species and adaptation to different ecological environments and breeding methodologies: environmentally friendly, resource-saving, product-diversified, usage-flexible, and more efficient breeding	perennial cereals (rice, wheat, and maize); diploid potato suitable for hybrid breeding; *de novo* domestication of wild plants; introgression of new alleles from closely related species

Table 4. Strategies for crop redesign at the micro and macro scales.
Revised from Zhang et al. (2021b).

topology, and clustering, and serves as a source of variation for novel traits (Hetti-Arachchilage et al., 2022). Predictive modeling and quantitative characterization of synthetic modules provide a detailed understanding of complex regulatory and signaling networks.

The genes, metabolic pathways and networks that control agronomic traits can be classified into independent modules, such as those for yield, quality, and resistance. Breeding by molecular module design began with the dissection of molecular modules for complex traits to identify allelic variations and gene network interactions (Xue et al., 2015). Molecular modules can be dissected into multiple components for a group of associated traits and then integrated to interact with one another through high-efficiency design and assembly. Finally, improved varieties can be developed through designed and predictive breeding by incorporating the modules and associated genes and networks in predictions. For example, using Kongyu131 as a base variety, several improved versions were developed by introgression of different trait modules, including a single-point substitution line with significant heading date delay (Wang et al., 2019a).

Future efforts will need to incorporate multiple layers of information to predict systems-level behavior of crop plant networks and their dynamics in changing environments. Network rewiring arises from changes in node and edge linkages, topological properties of individual nodes, subnetwork properties, and global topological properties. Hetti-Arachchilage et al. (2022) summarized several approaches for network rewiring: plant growth and stress responses that occur through signaling cascades can be fine-tuned by miRNA-mediated stress regulatory networks; gene regulatory networks can be mediated by heritable and robust epigenetic regulators; expression of regulatory and downstream stress-related genes can be fine-tuned by alternative splicing; diverse and complex protein translational modifications can be performed by dynamic control of the proteome; and natural and induced allelic variation can be leveraged for gene regulation.

Breeding by crop redesign at the macro scale

Crop redesign can also be performed at the macro scale for individuals, populations, and species (Table 4). Individual design dates back to the development of wheat and rice ideotypes by combining semi-dwarfism, erect top leaves, and strong stems (Jennings, 1964; Beachell and Jennings, 1965; Donald, 1968; Peng et al., 2008). The maize ideotype has been characterized by Mock and Pearce (1975), including stiff, vertically oriented leaves above the ear, a short interval between pollen shed and silk emergence, and small tassel size. At the population level, maize has been improved over the past decades to increase its suitability for high-density planting (from 30K to 79K plants/ha) (Duvick et al., 2004), contributing to significant yield increases. An array of prospective redesigns of plant systems, including straightforward alterations and conceptual redesigns, have been explored for improved photosynthetic efficiency and performance (Ort et al., 2015). Species design at the macro scale would dramatically reshape a plant species to adapt to a completely different ecological environment or production system (Tian et al., 2021). Two examples of such species design are perennial rice and seed-propagated diploid potato.

To increase food and ecosystem security, production of perennial grains has been proposed (Glover et al., 2010). To develop perennial crops, breeding programs must combine multiple desirable traits, including reliable regrowth with high yield and quality over multiple years, adaptation to abiotic stresses, and resistance to pests and diseases. A successful example of such population design is perennial varieties of rice. The first variety of rice capable of surviving for consecutive years, PR23, was bred in 2018 via clonal propagation of the rhizome from *Oryza longistaminata*. This variety was released in China and is now being tested throughout Asia and Africa (Huang et al., 2018).

Developing a self-pollinated diploid potato that can be bred through regular hybridization and selection is another example of crop redesign at the macro scale (here, species), although it will be achieved by genetic modification of many genes and pathways at the micro scale. Transforming potato from clonally propagated into seed propagated represents a significant innovation in plant breeding. Owing to deleterious mutations, it has been challenging to develop highly homozygous inbred potato lines. Redesign at the macro level was used to develop a generation of pure and fertile potato lines and thereby uniform, vigorous F1s. The redesign involved: (1) selection of starting materials, (2) genetic analysis of *Solanum* populations, (3) development of inbred lines, and (4) generation of vigorous F1 hybrids. During breeding, beneficial alleles were combined in the hybrids, and large-effect deleterious mutations were eliminated (Zhang et al., 2021a). The selection criteria included genome homozygosity and deleterious mutations in starting materials, segregation distortion in the S1 population, haplotype information for inferring the break of tight linkage between beneficial and deleterious alleles, and parental genome complementarity. Chromosome-scale and haplotype-resolved genome assembly enabled reconstruction of the four haplotypes of cultivated potato (Sun et al., 2022), ultimately increasing breeding success for the seed-propagated diploid.

Synthetic biology and *de novo* domestication

Metabolic pathways that are more efficient than native ones could be synthesized based on predictive results. An example is the attempt to install synthetic glycolate metabolic pathways in tobacco, maximizing flux through the pathways by inhibiting glycolate export from the chloroplast (South et al., 2019). With the synthetic pathways, photosynthetic quantum yield was improved by 20%, indicating that engineering alternative glycolate metabolic pathways into crop chloroplasts can drive increases in C3 crop yield. In rice, it was demonstrated that a partial C4 pathway can be established by transformation with a single construct harboring coding sequences for five enzymes of C4 metabolism (Ermakova et al., 2021). Although the enzyme expression levels require improvement, their cell-specific expression patterns were largely appropriate for two-cell C4 photosynthesis; furthermore, the observed photosynthetic phenotypes of the transgenic plants were consistent with the occurrence of C4 carboxylation *in vivo*, suggesting that a full C4 metabolic pathway may be possible in rice.

Synthetic modification has been used to develop nitrogen fixation capacities in nonlegume plants via rhizobial symbiosis

(Bozsoki et al., 2020). Plants evolved lysine motif (LysM) receptors to recognize and parse microbial elicitors and drive intracellular signaling to regulate microbial colonization. Two motifs in the LysM1 domains of chitin and nodulation (Nod) factor receptors of *Lotus japonicus* determine the recognition of specific ligands and enable discrimination between their *in planta* functions. Binding specificities in LysM receptors can be altered to facilitate Nod factor recognition and symbiotic signaling from a chitin receptor. A more recent study revealed the mechanism of nitrogen fixation in legume plants (Dong et al., 2021): a SHORTROOT–SCARECROW (SHR–SCR) stem cell program in cortical cells of the legume *Medicago truncatula* was found to specify their distinct fate. Legume species have a conserved cortical SHR–SCR network for response to rhizobial signals and initiation of cortical cell division for *de novo* nodule organogenesis.

Discovery of key domestication genes suggests that *de novo* domestication is a feasible method of crop redesign (Fernie and Yan, 2019; Tian et al., 2021; Chen et al., 2022c; Yu and Li, 2022). A new species could be shaped through selection of specific organs (Osnas, 2012), *de novo* domestication of a wild relative with different ploidy (Yu et al., 2021a), allopolyploidy among subspecies (Griffiths et al., 1999), or absorption of new alleles through introgression from nearby species (Arnold et al., 2016). Many genes have been studied for domestication traits, including fruit/seed size and weight, grain filling, growth habits, plant/inflorescence architecture, seed casing/color/dormancy, shattering, style length, and flowering (reviewed by Fernie and Yan, 2019). Wild crop relatives can be redomesticated to breed improved cultivated species that are adapted to climate change and environmental stresses. Traditional breeding approaches and gene editing have been proposed as two parallel approaches for *de novo* domestication (Fernie and Yan, 2019; Yu et al., 2021a). At present, gene editing in all plants relies on transformation technology. New technologies for addressing this dependency include high-efficiency transformation (Lowe et al., 2016), pollen-based transformation (Zhao et al., 2017), functional cloning of genes that overcomes genotype dependency (Wang et al., 2022a), and marker-assisted transfer of gene-edited traits.

TURNING SMART BREEDING INTO GENETIC GAIN

A comprehensive survey of US and European executives involved in big data initiatives identified the top 10 challenges in operationalizing big data and turning them into value (Capgemini Inc., 2016): IT budget constraints, data security concerns, integration challenges, lack of technical expertise, proliferation of data silos, corporate culture, poor or insufficient data quality, compliance concerns, user adoption and training needs, and manual processing/time constraints. In the field of plant breeding, big data and AI could potentially transform plant breeding from an art to a more data-driven science. Turning smart breeding into genetic gain and finally into breeding value and economic gain depends on how well all resources can be integrated and utilized together, including big data, AI, and innovative breeding technologies such as iGEP. In developing countries and small- and medium-sized breeding enterprises, sharing breeding platforms and resources while protecting the intellectual property of participants will be critical for the application of smart breeding.

Turning plant breeding from art to more data-driven science

Largely owing to limited information and resources for trait prediction, plant breeding has long been considered both an art and a science. However, prediction models can be constructed and validated using real data, simulated data, and their combination. This procedure can be facilitated by AI. In the era of smart breeding, breeders can be both generators and users of big data. We are moving into a new stage of mega data- and cloud technology-assisted breeding. Transition of plant breeding from an art to a data-driven science will have four significant features: targeted design, predicted selection, pipeline-driven approaches, and step-by-step improvement by "standing on the giant's shoulder" (using the best base variety or material for further improvement). Marker-assisted plant breeding requires many years in developing countries and small- and medium-sized enterprises (Xu and Crouch, 2008). Smart breeding is expected to take approximately the same amount of time to become a routine practice worldwide. Even if it becomes widely applied in some countries or MSEs, it will take years to spread out and be recognized by the whole community, similar to the application of GS from livestock to crop plants (Xu et al., 2020a).

Smart breeding driven by big data, AI, and iGEP can use all available information and resources collected spatiotemporally. With optimized models generated using well-selected data and populations, prediction should be equivalent to a comprehensive evaluation across multiple years and MET sites. Therefore, predictive breeding can reveal phenotypic performance better than any single observation. Although breeders may place greater trust in what they can see with the naked eye in the field, each such observation is just one of many samplings required for accurate judgement. Evaluation and thus selection from such a random sample will vary greatly with time, location, and individual breeders.

Modern plant breeding has become a field with a great deal of technological support and a breeding pipeline that brings together breeders, technicians, platforms, facilities, experimental stations, and METs. For example, an intelligent greenhouse has many components, including monitors, sensors, and controllers. Smart breeding should aim to integrate the "internet of things" with intelligent facilities and crop management, driven by big data, AI, and iGEP (Figure 3).

Integrative breeding programs

Although application of new breeding technologies has been largely crop dependent, most crop species share a similar panel of innovations (Figure 3). Breeding facilities (e.g., genotyping, phenotyping, and envirotyping), information management (breeding data archiving, integration, analytics, and mining, through a system such as Tripal; Staton et al., 2021), and decision support (simulation, prediction, validation, optimization, and selection) can be integrated, and improved germplasm and varieties can be protected through big-data-assisted plant variety protection (Figure 1).

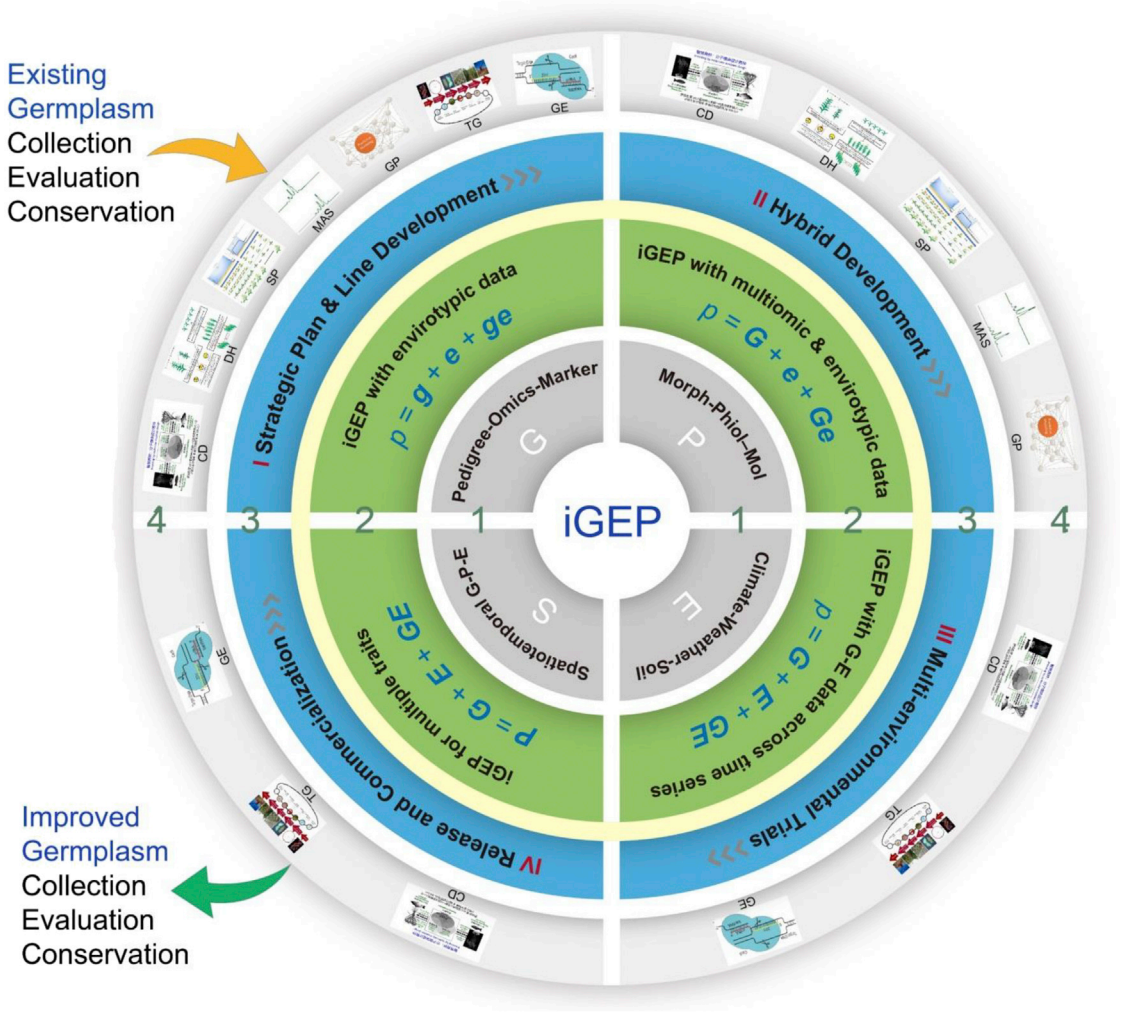

Figure 3. Conceptualization of a smart breeding platform driven by big data, artificial intelligence, and iGEP.
Beginning at the center, level 1 represents big data. Moving outward, level 2 represents iGEP models (S, spatiotemporal G-P-E). In the formulas, lowercase, bold lowercase, and bold uppercase letters represent single, vector, and matrix variables, respectively. Level 3 shows a breeding pipeline from strategic plan and line development through hybrid development and multi-environmental trials to release and commercialization. Level 4 shows new breeding technologies (CD, crop design; DH, doubled haploids; SP, speed breeding; MAS, marker-assisted selection; GP, genomic prediction; TG, transgenics; GE, genome editing). The white cross that divides the figure into four quadrants represents information flow between layers and circles, integrating all components together like an internet of things. The large arrows in circle 3 represent the direction of the breeding workflow and pipeline, starting from I, moving to IV, and circling back through the breeding pipeline. Divided by the yellow circle, the inside contains levels 1 and 2 that share data and models, and the outside contains levels 3 and 4 that share the breeding pipeline and new breeding technologies.

Although genetic transformation has been used in commercialized varieties globally, transgenic technology remains to be fully integrated into routine breeding pipelines. One of the major limitations is that most base varieties are recalcitrant to genetic transformation. The same situation occurs with genome editing using CRISPR/Cas9 (Jinek et al., 2012), in which genetic transformation is also required. Once transformations are completed using a tissue-culture-friendly variety, marker-assisted backcrossing (MAB) can then be used to introgress target genes from transgenic or genome-edited lines into base varieties or desirable breeding materials. MAB has been widely used in MSEs to simultaneously introgress two to seven transgenes into several hundreds of inbreds or varieties each year (Kunsheng Wu 2017, personal communication).

Plant breeding programs can be accelerated by shortening the number of generations required to reach pure breeding and by speeding plant growth and development (Xu et al., 2017; Sinha et al., 2021). DH approaches can generate true-breeding lines in two generations (De La Fuente et al., 2013), compared with the eight or more generations required with continuous inbreeding, significantly shortening the breeding cycle. The gene that controls haploid induction in maize has been cloned (Kelliher et al., 2017; Liu et al., 2017), and DH breeding is now expected to be used for all crop species through genomic editing (e.g., Wang et al., 2022b). Although pollen culture has been very successful in the generation of DH lines, DH breeding has not been widely used in self-pollinated crop species, largely because generation of true-breeding lines by selfing is much easier than with outcrossed species. It is expected

that combining cloned genes with genome editing (Yao et al., 2018; Liu et al., 2021) will stimulate DH breeding in self-pollinated crops.

In contrast to DH-based approaches, speed breeding can be used to accelerate the generation time and thus shorten the breeding cycle for all crop species and breeding programs. However, the environments required for the speed breeding process will be crop specific and case dependent. Combining embryo culture with management of water stress, light intensity and duration, temperature, and potting mixture enables the production of up to eight generations of wheat and nine generations of barley per year (Zheng et al., 2013). For some crops (wheat, barley, chickpea, and canola), growth in a managed environment with an extended photoperiod (22/2 h light/dark) causes plants to flower in a much shorter time (Watson et al., 2018). Adjustment of photoperiod and temperature can be combined with alterations in other factors that affect growth and flowering, such as nutrients and microelement levels. Reproduction processes could be accelerated by regulating all of the genes that affect plant growth and development (Fasong Zhou, personal communication). MAS can be used to identify the best combination of relevant genes for manipulation. As a result, a suite of accelerating factors could be developed for each crop species or for specific genotypes.

In response to the challenge of GS with big data, Bayer built an AI assistant that helps breeders select the right candidates. Cloud-based algorithms built on a foundation of roughly 1.7 trillion calculations enables a dramatic shift in the scale and speed of the breeding pipeline. Predicting the potential of multiple breeding lines allows breeders to disregard certain lines that are not likely to achieve the intended goal. With neural networks, AI-assisted models are literally learning throughout the entire process, providing breeders a road map to follow with enhanced accuracy and efficiency (https://www.bayer.com/en/agriculture/article/how-math-and-data-science-accelerate-innovation-while-conserving).

Open-source breeding

Large-scale commercial breeding programs typically operate as a coordinated network, increasing the efficiency of breeding platforms and saving costs (Xu et al., 2020a). These programs generate enough big data to build their own iGEP models. With open-source breeding, small- and medium-sized breeding enterprises will each function as a breeding team within an MSE. Supported by synthetic pipeline services, low-cost genotyping platforms, and the capacity to obtain models and haplotype effects, partners can share their breeding data and materials after each breeding cycle. This will allow each member to move to the next cycle with accumulated information and materials, updated designs, and optimized models, ultimately improving genetic gain (Xu et al., 2017, 2020a). Such an open-source breeding initiative creates additional challenges with respect to big data management, model construction, and prediction, owing to the increased complexity of the breeding information being incorporated. In open-source breeding programs, the intellectual property of plant varieties can be protected using a highly efficient, low-cost molecular marker system such as genotyping by target sequencing and liquid chip (Guo et al., 2021). As DNA profiles can be constructed for almost all selections and breeding materials that have been or are being shared in the initiative, an acceptable royalty system can be established as a revenue generator.

The prediction accuracy of ML models is largely determined by the quality of the dataset employed (LeCun et al., 2015). Consequently, a frequent challenge for training robust ML models is the lack of appropriate datasets with enough data points and sample variability. A few international consortia, such as AgBioData (Harper et al., 2018) and Breeding API (Selby et al., 2019), are making an effort to share and transform breeding datasets. However, a centralized platform for collecting, hosting, and managing enviromic data is needed to make the relevant data more widely available, similar to the approaches used to share other omics data. An alternative approach to protect sensitive information while supporting collaboration toward data-driven breeding is the establishment of federated learning cohorts (Konečný et al., 2016). Within these, each participant institution trains the model with its own dataset and shares the updated model peer to peer or through a centralized server that will aggregate the model's weights. The updated model parameters thus improve the baseline model, which is then shared among institutions (Yang et al., 2019a, 2019b). Another factor that prevents researchers from employing previously published datasets is the lack of standardized metadata descriptions, including experimental design, data collection protocol, field management, environmental variables, and other information (Danilevicz et al., 2022). The minimum information about a plant phenotyping project (Papoutsoglou et al., 2020) offers a resource to guide researchers in annotating metadata to increase usability and interoperability.

There are several ongoing open-source breeding programs in China, including maize molecular breeding initiatives and integrated plant breeding platforms, members of which share nearly all of their data and materials (Xu et al., 2020a). By integrating the breeding technologies discussed above (Figure 3), open-source breeding initiatives can be established in locations with nontarget environments. An example is Hainan, China, where almost all breeding enterprises perform off-season breeding (Zhang et al., 2021b). Using 44 624 wheat lines and over 7.6 million genotyping-by-sequencing data points, a reference wheat genotype–phenotype map was built with a large number of marker-trait associations (Juliana et al., 2019), providing a valuable resource for worldwide open-source breeding in wheat.

International initiatives, such as the Excellence in Breeding Platform (http://excellenceinbreeding.org), enable greater technology uptake by breeders and farmers and integrate the community globally. In addition to the five modules established, envirotyping should also be included as an independent module. With a well-established open-source breeding system, the accumulation of small dispersed datasets from all partners and collaborators builds large and diverse datasets that can be shared, "many a little making a mickle." All public information, including genotypic data for germplasm bank materials and envirotypic data for weather, climate, and soil, should be made available for open-source breeding.

CONCLUDING REMARKS

Future plant breeding will become smart with the implementation of big data, AI, and iGEP. Smart breeding will enable all breeding-related information, including spatiotemporal omics, to be broadly accessible in a usable format. Innovative breeding technologies, including transgenics, genome editing, DH-based breeding, speed breeding, and MAS, will be integrated into a smart breeding pipeline (Figure 3). Model selection and optimization can be performed in iGEP, with the best option identified for each breeding population. Incorporation of enviromics can improve prediction accuracy through better understanding of GEIs, enhanced phenotyping precision, managed environments, and optimized model construction. Establishment of integrative plant breeding platforms and open-source breeding initiatives will aid in translating smart breeding efforts into genetic gain in the breeding pipelines of national programs as well as small- and medium-sized breeding enterprises.

FUNDING

This research is supported by the National Key Research and Development Program of China (2016YFD0101803), the Central Public-interest Scientific Institution Basal Research Fund (Y2020PT20), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences (CAAS-XTCX2016009), the Shijiazhuang Science and Technology Incubation Program (191540089A), the Hebei Innovation Capability Enhancement Project (19962911D), the Project of Hainan Yazhou Bay Seed Laboratory (B21HJ0223), and the Department of Science and Technology of Ninxia Project (NXNYYZ202001). Research activities at CIMMYT were supported by the Bill and Melinda Gates Foundation and the CGIAR Research Program MAIZE.

ACKNOWLEDGMENTS

The authors thank Drs. Jianming Yu (Iowa State University, USA), Xing Wang Deng (Peking University Institute of Advanced Agricultural Sciences, China), Hongwei Zhang (Chinese Academy of Agricultural Sciences, China), and three anonymous reviewers for their critical reading and constructive comments on the manuscript. No conflict of interest is declared.

Received: April 4, 2022
Revised: August 20, 2022
Accepted: September 2, 2022
Published: September 7, 2022

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol **52**:12.

Acosta Pech, R., Crossa, J., de los Campos, G., Teyssèdre, S., Claustres, B., Pérez-Elizalde, S., and Pérez-Rodríguez, P. (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor. Appl. Genet **130**:1431–1440.

An, B., Gao, X., Chang, T., Xia, J., Wang, X., Miao, J., Xu, L., Zhang, L., Chen, Y., Li, J., et al. (2020). Genome-wide association studies using binned genotypes. Heredity **124**:288–298.

Ansarifar, J., Akhavizadegan, F., and Wang, L. (2020). Performance prediction of crosses in plant breeding through genotype by environment interactions. Sci. Rep. **10**:11533.

Araus, J.L., Kefauver, S.C., Zaman-Allah, M., Olsen, M.S., and Cairns, J.E. (2018). Translating high-throughput phenotyping into genetic gain. Trends Plant Sci. **23**:451–466.

Arnold, B.J., Lahner, B., DaCosta, J.M., Weisman, C.M., Hollister, J.D., Salt, D.E., Bomblies, K., and Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. Proc. Natl. Acad. Sci. USA **113**:8320–8325.

Atkinson, J.A., Pound, M.P., Bennett, M.J., and Wells, D.M. (2019). Uncovering the hidden half of plants using new advances in root phenotyping. Curr. Opin. Biotechnol. **55**:1–8.

Auinger, H.-J., Lehermeier, C., Gianola, D., Mayer, M., Melchinger, A.E., da Silva, S., Knaak, C., Ouzunova, M., and Schön, C.C. (2021). Calibration and validation of predicted genomic breeding values in an advanced cycle maize population. Theor. Appl. Genet. **134**:3069–3081.

Azodi, C.B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., and Shiu, S.H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 (Bethesda) **9**:3691–3702.

Azodi, C.B., Tang, J., and Shiu, S.-H. (2020). Opening the black box: interpretable machine learning for geneticists. Trends Genet. **36**:442–455.

Baker, R.J. (1986). Selection Indices in Plant Breeding (CRC Press).

Banerjee, R., Marathi, B., and Singh, M. (2020). Efficient genomic selection using ensemble learning and ensemble feature reduction. J. Crop Sci. Biotechnol. **23**:311–323.

Beachell, H.M., and Jennings, P.R. (1965). Need for modification of plant type. In The Mineral Nutrition of the Rice Plant (Baltimore: John Hopkins Press), pp. 29–35.

Beans, C. (2020). Crop researchers harness artificial intelligence to breed crops for the changing climate. Proc. Natl. Acad. Sci. USA **117**:27066–27069.

Beckers, J., Wurst, W., and de Angelis, M.H. (2009). Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling. Nat. Rev. Genet. **10**:371–380.

Bellman, R.E. (1961). Adaptive Control Processes (Princeton, NJ: Princeton University Press).

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. **34**:20–25.

Bernardo, R. (2016). Bandwagons I, too, have known. Theor. Appl. Genet. **129**:2323–2332.

Bernardo, R. (2021). Predictive breeding in maize during the last 90 years. Crop Sci. **61**:2872–2881.

Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. Crop Sci. **47**:1082–1090.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., et al. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci. **55**:154–163.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**:111–135.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., and Lang, M. (2020). Benchmark for filter methods for feature selection in highdimensional classification data. Comput. Stat. Data Anal. **143**:106839.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell **132**:311–322.

Bozsoki, Z., Gysel, K., Hansen, S.B., Lironi, D., Krönauer, C., Feng, F., de Jong, N., Vinther, M., Kamble, M., Thygesen, M.B., et al. (2020).

Ligand-recognizing motifs in plant LysM receptors are major determinants of specificity. Science **369**:663–670.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science **296**:752–755.

Budhlakoti, N., Mishra, D.C., Rai, A., Lal, S.B., Chaturvedi, K.K., and Kumar, R.R. (2019). A comparative study of single-trait and multi-trait genomic selection. J. Comput. Biol. **26**:1100–1112.

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. **52**:707–719.

Capgemini Inc. (2016). Big data payoff-turning big data into business value. https://www.capgemini.com/.

Chan, M. (2018). Big data in the cloud: why cloud computing is the answer to your big data initiatives. https://www.thorntech.com/big-data-in-the-cloud/.

Che, D., Liu, Q., Rasheed, K., and Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv. Exp. Med. Biol. **696**:191–199.

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.

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. (2022a). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell **185**:1777–1792.e21.

Chen, R., Deng, Y., Ding, Y., Guo, J., Qiu, J., Wang, B., Wang, C., Xie, Y., Zhang, Z., Chen, J., et al. (2022b). Rice functional genomics: decades' efforts and roads ahead. Sci. China Life Sci. **65**:33–92.

Chen, W., Chen, L., Zhang, X., Yang, N., Guo, J., Wang, M., Ji, S., Zhao, X., Yin, P., Cai, L., et al. (2022c). Convergent selection of a WD40 protein that enhances grain yield in maize and rice. Science **375**:eabg7985.

Chen, C.J., Rutkoski, J., Schnable, J.C., Murray, S.C., Wang, L., Jin, X., and Stich, B. (2023). Role of the genomics–phenomics–agronomy paradigm in plant breeding. Plant Breed. Rev. **46**:622–667.

Cheng, C.Y., Li, Y., Varala, K., Bubert, J., Huang, J., Kim, G.J., Halim, J., Arp, J., Shih, H.J.S., Levinson, G., et al. (2021). Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat. Commun. **12**:5627.

Chollet, F., and Allaire, J.J. (2017). Deep Learning with R. Manning Publications, Manning Early Access Program (MEA), 1st edn.

Cooper, M., and Messina, C.D. (2021). Can we harness "enviromics" to accelerate crop improvement by integrating breeding and agronomy? Front. Plant Sci. **12**:735143.

Cooper, M., Gho, C., Leafgren, R., Tang, T., and Messina, C. (2014a). Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. J. Exp. Bot. **65**:6191–6204.

Cooper, M., Messina, C.D., Podlich, D., Totir, L.R., Baumgarten, A., Hausmann, N.J., Wright, D., and Graham, G. (2014b). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci. **65**:311–336.

Cooper, M., Technow, F., Messina, C., Gho, C., and Totir, L.R. (2016). Use of crop growth models with whole-genome prediction: application to a maize multi-environment trial. Crop Sci. **56**:2141–2156.

Cooper, M., Tang, T., Gho, C., Hart, T., Hammer, G., and Messina, C. (2020). Integrating genetic gain and gap analysis to predict improvements in crop productivity. Crop Sci. **60**:582–604.

Coppens, F., Wuyts, N., Inzé, D., and Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. Curr. Opin. Struct. Biol. **4**:58–63.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity **126**:92–106.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity **126**:92–106.

Costa-Neto, G., Galli, G., Carvalho, H.F., Crossa, J., and Fritsche-Neto, R. (2021b). EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. G3 **11**:jkab040.

Cox, M., and Ellsworth, D. (1997). Managing big data for scientific visualization. ACM Siggraph **97**:21–38.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. **22**:961–975.

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O.A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., and Bentley, A.R. (2021). The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. Front. Plant Sci. **12**:651480.

Daniel, R. (2005). The metagenomics of soil. Nat. Rev. Microbiol. **3**:470–478.

Danilevicz, M.F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P.E., and Edwards, D. (2022). Plant genotype to phenotype prediction using machine learning. Front. Genet. **13**:822173.

De La Fuente, G.N., Frei, U.K., and Lübberstedt, T. (2013). Accelerating plant breeding. Trends Plant Sci. **18**:667–672.

Denison, R.F. (2015). Evolutionary tradeoffs as opportunities to improve yield potential. Field Crop. Res. **182**:3–8.

De los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**:327–345.

De Sousa, K., van Etten, J., Poland, J., Fadda, C., Jannink, J.-L., Kidane, Y.G., Lakew, B.F., Mengistu, D.K., Pè, M.E., Solberg, S.Ø., et al. (2021). Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. Commun. Biol. **4**:944.

Diepenbrock, C.H., Tang, T., Jines, M., Technow, F., Lira, S., Podlich, D., Cooper, M., and Messina, C. (2022). Can we harness digital technologies and physiology to hasten genetic gain in US maize breeding? Plant Physiol. **188**:1141–1157. kiab527.

Donald, C.M. (1968). The breeding of crop ideotypes. Euphytica **17**:385–403.

Dong, W., Zhu, Y., Chang, H., Wang, C., Yang, J., Shi, J., Gao, J., Yang, W., Lan, L., Wang, Y., et al. (2021). An SHR–SCR module specifies legume cortical cell fate to enable nodulation. Nature **589**:586–590.

Doxtator, C.W., and Johnson, I.J. (1936). Prediction of double cross yields in corn. Agron. J. **28**:460–462.

Duvick, D.N. (2005). The contribution of breeding to yield advances in maize (*Zea mays* L.). Adv. Agron. **86**:83–145.

Duvick, D.N., Smith, J.S.C., and Cooper, M. (2004). Long-term selection in a commercial hybrid maize breeding program. Plant Breed. Rev. **24**:109–151.

Dy, J.G., Brodley, C.E., Kak, A., Broderick, L.S., and Aisen, A.M. (2003). Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Trans. Pattern Anal. Mach. Intell. **25**:373–378.

Eathington, S.R., Crosbie, T.M., Edwards, M.D., Reiter, R.S., and Bull, J.K. (2007). Molecular markers in a commercial breeding program. Crop Sci. **47**:S-154–S-163.

Edwards, M., and Johnson, L. (1994). RFLPs for rapid recurrent selection. In Analysis of Molecular Marker Data (ASHS and CSSA), pp. 33–40.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**:e19379.

Ermakova, M., Arrivault, S., Giuliani, R., Danila, F., Alonso-Cantabrana, H., Vlad, D., Ishihara, H., Feil, R., Guenther, M., Borghi, G.L., et al. (2021). Installation of C4 photosynthetic pathway enzymes in rice using a single construct. Plant Biotechnol. J. **19**:575–588.

Fan, J., and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the International Congress of Mathematicians (Madrid, Spain: European Mathematical Society). p595-622.

Fernie, A.R., and Yan, J. (2019). De novo domestication: an alternative route toward new crops for the Future. Mol. Plant **12**:615–631.

Flores, F., Moreno, M.T., and Cubero, J.I. (1998). A comparison of univariate and multivariate methods to analyze G X E interaction. Field Crop. Res. **56**:271–286.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. **3**:1289–1305.

Fu, J., Hao, Y., Li, H., Reif, J.C., Chen, S., Huang, C., Wang, G., Li, X., Xu, Y., and Li, L. (2022). Integration of genomic selection with doubled-haploid evaluation in hybrid breeding from GS1.0 to GS4.0 and beyond. Mol. Plant **15**:577–580.

Gabur, I., Simioniuc, D.P., Snowdon, R.J., and Cristea, D. (2022). Machine learning applied to the search for nonlinear features in breeding populations. Front. Artif. Intell. **5**:876578.

Gärtner, T., Steinfath, M., Andorf, S., Lisec, J., Meyer, R.C., Altmann, T., Willmitzer, L., and Selbig, J. (2009). Improved heterosis prediction by combining information on DNA- and metabolic markers. PLoS One **4**:e5220.

Gil, Y., Greaves, M., Hendler, J., and Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. Science **346**:171–172.

Gill, M., Anderson, R., Hu, H., Bennamoun, M., Petereit, J., Valliyodan, B., Nguyen, H.T., Batley, J., Bayer, P.E., and Edwards, D. (2022). Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC Plant Biol. **22**:180.

Glover, J.D., Reganold, J.P., Bell, L.W., Borevitz, J., Brummer, E.C., Buckler, E.S., Cox, C.M., Cox, T.S., Crews, T.E., Culman, S.W., et al. (2010). Increased food and ecosystem security via perennial grains. Science **328**:1638–1639.

Gosa, S.C., Lupo, Y., and Moshelion, M. (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.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., and Lewontin, R.C. (1999). Modern Genetic Analysis (New York: W.H. Freeman).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (Cambridge: MAMIT Press).

Grinberg, N.F., Orhobor, O.I., and King, R.D. (2019). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice and wheat. Mach. Learn. **109**:251–277.

Guo, T., Mu, Q., Wang, J., Vanous, A.E., Onogi, A., Iwata, H., Li, X., and Yu, J. (2020). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. Genome Res. **30**:673–683.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor. Appl. Genet. **129**:2413–2427.

Guo, Z., Yang, Q., Huang, F., Zheng, H., Sang, Z., Xu, Y., Zhang, C., Wu, K., Tao, J., Prasanna, B.M., et al. (2021). Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. Plant Commun. **2**:100230.

Guyon, I., and Elissee, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res. **3**:1157–1182.

Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). Extension of the bayesian alphabet for genomic selection. BMC Bioinf. **12**:186.

Harbinson, J., Parry, M.A.J., Davies, J., Rolland, N., Loreto, F., Wilhelm, R., Metzlaff, K., and Klein Lankhorst, R. (2021). Designing the crops for the future. Biology **10**:690.

Harfouche, A.L., Jacobson, D.A., Kainer, D., Romero, J.C., Harfouche, A.H., Scarascia Mugnozza, G., Moshelion, M., Tuskan, G.A., Keurentjes, J.J.B., and Altman, A. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol. **37**:1217–1235.

Harper, L., Campbell, J., Cannon, E.K.S., Jung, S., Poelchau, M., Walls, R., Andorf, C., Arnaud, E., Berardini, T.Z., Birkett, C., et al. (2018). AgBioData Consortium recommendations for sustainable genomics and genetics databases for agriculture. Database **2018**:bay088.

Heffner, E.L., Sorrells, M.E., and Jannink, J.L. (2009). Genomic selection for crop improvement. Crop Sci. **49**:1–12.

Heslot, N., Yang, H.-P., Sorrells, M.E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. Crop Sci. **52**:146–160.

Heslot, N., Akdemir, D., Sorrells, M.E., and Jannink, J.L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. **127**:463–480.

Hetti-Arachchilage, M., Challa, G.S., and Marshall-Colón, A. (2022). Rewiring network plasticity to improve crops. Plant Breed. Rev. **45**:143–183.

Hey, T., Stewart Tansley, S., and Tolle, K. (2009). The Fourth Paradigm: Data Intensive Scientific Discovery (Microsoft Research)978-0-9825442-0-4.

Hospital, F., Moreau, L., Lacoudre, F., Charcosset, A., and Gallais, A. (1997). More on the efficiency of marker-assisted selection. Theor. Appl. Genet. **95**:1181–1189.

Hu, X., Xie, W., Wu, C., and Xu, S. (2019). A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol. J. **17**:2011–2020.

Huang, G., Qin, S., Zhang, S., Cai, X., Wu, S., Dao, J., Zhang, J., Huang, L., Harnpichitvitaya, D., Wade, L., et al. (2018). Performance, economics and potential impact of perennial rice PR23 relative to annual rice cultivars at multiple locations in Yunnan province of China. Sustainability **10**:1086.

Ibba, M.I., Crossa, J., Montesinos-López, O.A., Montesinos-López, A., Juliana, P., Guzman, C., Delorean, E., Dreisigacker, S., and

Poland, J. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. Plant Genome **13**:e20034.

Jansen, R.C., and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. Trends Genet. **17**:388–391.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. **127**:595–607.

Jenkins, M.T. (1934). Methods of estimating the performance of double crosses in corn. Agron. J. **26**:199–204.

Jennings, P.R. (1964). Plant type as a rice breeding objective. Crop Sci. **4**:13–15.

Jia, Y., and Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics **192**:1513–1522.

Jiang, J., Zhang, Q., Ma, L., Li, J., Wang, Z., and Liu, J.F. (2015). Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. Heredity **115**:29–36.

Jin, X., Zarco-Tejada, P.J., Schmidhalter, U., Reynolds, M.P., Hawkesford, M.J., Varshney, R.K., Yang, T., Nie, C., Li, Z., Ming, B., et al. (2021). High-throughput estimation of crop traits: a review of ground and aerial phenotyping platforms. IEEE Geosci. Remote Sens. Mag. **9**:200–231.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science **337**:816–821.

Johnson, W.B., and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a Hilbert space. In Conference in Modern Analysis and Probability (New Haven, Conn., 1982). Contemporary Mathematics. 26. Providence, RI (American Mathematical Society), pp. 189–206.

Jubair, S., and Domaratzki, M. (2019). Ensemble supervised learning for genomic selection. In In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), pp. 1993–2000.

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F.H., Govindan, V., Mondal, S., Kumar, U., et al. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat. Genet. **51**:1530–1539.

Kang, M.S. (2002). Genotype-environment interaction: progress and prospects. In Quantitative Genetics, Genomics and Plant Breeding, M.S. Kang, ed. (CAB International), pp. 221–243.

Kelliher, T., Starr, D., Richbourg, L., Chintamanani, S., Delzer, B., Nuccio, M.L., Green, J., Chen, Z., McCuiston, J., Wang, W., et al. (2017). MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. Nature **542**:105–109.

Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. Front. Plant Sci. **10**:621.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2020). Predicting yield performance of parents in plant breeding: a neural collaborative filtering approach. PLoS One **15**:e0233382.

Kim, K.H., Kim, J.Y., Lim, W.J., Jeong, S., Lee, H.Y., Cho, Y., Moon, J.K., and Kim, N. (2020). Genome-wide association and epistatic interactions of flowering time in soybean cultivar. PLoS One **15**:e0228114.

Koch, P., Wujek, B., Golovidov, O., and Gardner, S. (2017). Automated hyperparameter tuning for effective machine learning. In Proceedings of the SAS Global Forum 2017 Conference (Cary, NC: SAS Institute Inc). http://support.sas.com/resources/papers/proceedings17/SAS514-2017.pdf.

Konečný, J., McMahan, H.B., Ramage, D., and Richtárik, P. (2016). Federated optimization: distributed machine learning for on-device intelligence. Preprint at arXiv. CoRR abs/1610.02527.

Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling (New York: Springer).

Kusmec, A., Zheng, Z., Archontoulis, S., Ganapathysubramanian, B., Hu, G., Wang, L., Yu, J., and Schnable, P.S. (2021). Interdisciplinary strategies to enable data-driven plant breeding in a changing climate. One Earth **4**:372–383.

Kyratzis, A.C., Skarlatos, D.P., Menexes, G.C., Vamvakousis, V.F., and Katsiotis, A. (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.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**:743–756.

Lantz, B. (2015). Machine Learning with R, 2nd edn. (Birmingham: Packt Publishing Ltd).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature **521**:436–444.

Lee, T.G. (2021). Artificial intelligence for breeding better crops. https://vscnews.com/artificial-intelligence-crop-breeding/.

Lee, T.G. (2020). High-performance computing for breeding better crops. https://specialtycropindustry.com/high-performance-computing-for-breeding-better-crops/. [Accessed 2 October 2022].

Li, D., Quan, C., Song, Z., Li, X., Yu, G., Li, C., and Muhammad, A. (2021a). High-throughput plant phenotyping platform (HT3P) as a novel tool for estimating agronomic traits from the lab to the field. Front. Bioeng. Biotechnol. **8**:623705.

Li, X., Guo, T., Bai, G., Zhang, Z., See, D., Marshall, J., Garland-Campbell, K.A., and Yu, J. (2022a). Genetics-inspired data-driven approaches explain and predict crop performance fluctuations attributed to changing climatic conditions. Mol. Plant **15**:203–206.

Li, M., Zhang, Y.-W., Zhang, Z.-C., Xiang, Y., Liu, M.-H., Zhou, Y.-H., Zuo, J.-F., Zhang, H.-Q., Chen, Y., and Zhang, Y.-M. (2022b). A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol. Plant **15**:630–650.

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc. Natl. Acad. Sci. USA **115**:6679–6684.

Li, X., Guo, T., Wang, J., Bekele, W.A., Sukumaran, S., Vanous, A.E., McNellie, J.P., Tibbs-Cortes, L.E., Lopes, M.S., Lamkey, K.R., et al. (2021b). An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. Mol. Plant **14**:874–887.

Liu, C., Li, X., Meng, D., Zhong, Y., Chen, C., Dong, X., Xu, X., Chen, B., Li, W., Li, L., et al. (2017). A 4-bp insertion at ZmPLA1 encoding a putative phospholipase A generates haploid induction in maize. Mol. Plant **10**:520–522.

Liu, J., Liang, D., Yao, L., Zhang, Y., Liu, C., Liu, Y., Wang, Y., Zhou, H., Kelliher, T., Zhang, X., et al. (2021). Rice haploid inducer development by genome editing. Methods Mol. Biol. **2238**:221–230.

López, O.A.M., López, A.M., and Crossa, J. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction (Switzerland: Springer Nature Switzerland AG).

Lourenço, V.M., Ogutu, J.O., Rodrigues, R.A.P., and Piepho, H.-P. (2022). Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based

and deep learning methods on synthetic and empirical data. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.09.495423.

Lowe, K., Wu, E., Wang, N., Hoerster, G., Hastings, C., Cho, M.J., Scelonge, C., Lenderts, B., Chamberlin, M., Cushatt, J., et al. (2016). Morphogenic regulators baby boom and wuschel improve monocot transformation. Plant Cell **28**:1998–2015.

Lund, H. (2020). Most common types of data integration methods. https://www.rapidionline.com/blog/most-common-types-of-data-integration-methods.

Luo, L., Mei, H., Yu, X., Xia, H., Chen, L., Liu, H., Zhang, A., Xu, K., Wei, H., Liu, G., et al. (2019). Water-saving and drought-resistance rice: from the concept to practice and theory. Mol. Breed. **39**:145.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta **248**:1307–1318.

Marsh, J.I., Hu, H., Gill, M., Batley, J., and Edwards, D. (2021). Crop breeding for a changing climate: integrating phenomics and genomics with bioinformatics. Theor. Appl. Genet. **134**:1677–1690.

Marx, V. (2013). The big challenges of big data. Nature **498**:255–260.

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., Iwata, H., Li, Y., Lipka, A.F., and Zhang, Z. (2022). Ideas in genomic selection with the potential to transform plant molecular breeding: a review. Plant Breed. Rev. **45**:273–319.

Merrick, L.F., and Carter, A.H. (2021). Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. Plant Genome **14**:e20158.

Messina, C.D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: integrating crop growth models (CGM) with whole genome prediction (WGP). Eur. J. Agron. **100**:151–162.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**:1819–1829.

Millet, E.J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., Charcosset, A., Welcker, C., van Eeuwijk, F., and Tardieu, F. (2019). Genomic prediction of maize yield across European environmental conditions. Nat. Genet. **51**:952–956.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. Genes **10**:87.

Mock, J.J., and Pearce, R.B. (1975). An ideotype of maize. Euphytica **24**:613–623.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Toledo, F.H., Pérez-Hernández, O., Eskridge, K.M., and Rutkoski, J. (2016). A genomic Bayesian multi-trait and multi-environment model. G3 **6**:2725–2744.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C.M., and Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3 (Bethesda) **8**:3829–3840.

Montesinos-López, O.A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J.A., Martini, J.W.R., Fajardo-Flores, S.B., Gaytan-Lugo, L.S., Santana-Mancilla, P.C., and Crossa, J. (2021a). A review of deep learning applications for genomic selection. BMC Genom. **22**:19.

Montesinos-López, A., Runcie, D.E., Ibba, M.I., Pérez-Rodríguez, P., Montesinos-López, O.A., Crespo, L.A., Bentley, A.R., and Crossa, J. (2021b). Multi-trait genomic-enabled prediction enhances accuracy in multi-year wheat breeding trials. G3 **11**:kab270.

Montesinos-López, O.A., Montesinos-López, A., Hernandez-Suarez, C.M., Barrón-López, J.A., and Crossa, J. (2021c). Deep-learning power and perspectives for genomic selection. Plant Genome **14**:e20122.

Montesinos-López, O.A., Gonzalez, H.N., Montesinos-López, A., Daza-Torres, M., Lillemo, M., Montesinos-López, J.C., and Crossa, J. (2022). Comparing gradient boosting machine and Bayesian threshold BLUP for genome-based prediction of categorical traits in wheat breeding. Plant Genome **2022**:e20214.

Moore, B.M., Wang, P., Fan, P., Lee, A., Leong, B., Lou, Y.-R., Schenck, C.A., Sugimoto, K., Last, R., and Lehti-Shiu, M.D. (2020). Within-and cross-species predictions of plant specialized metabolism genes using transfer learning. in silico. Plants **2**:diaa005.

Moose, S.P., Dudley, J.W., and Rocheford, T.R. (2004). Maize selection passes the century mark: a unique resource for 21st century genomics. Trends Plant Sci. **9**:358–364.

Morais, R., Silva, N., Mendes, J., Adão, T., Pádua, L., López-Riquelme, J., Pavón-Pulido, N., Sousa, J.J., and Peres, E. (2019). mySense: a comprehensive data management environment to improve precision agriculture practices. Comput. Electron. Agric. **162**:882–894.

Morgan, L. (2018). What is data integration and how does it work?. https://www.datamation.com/big-data/what-is-data-integration/.

Morisse, M., Wells, D.M., Millet, E.J., Lillemo, M., Fahrner, S., Cellini, F., Lootens, P., Muller, O., Herrera, J.M., Bentley, A.R., et al. (2022). A European perspective on opportunities and demands for field-based crop phenotyping. Field Crop. Res. **276**:108371.

Munné-Bosch, S. (2022). Spatiotemporal limitations in plant biology research. Trends Plant Sci. **27**:346–354.

Muranty, H., Troggio, M., Sadok, I.B., Rifaï, M.A., Auwerkerken, A., Banchi, E., Velasco, R., Stevanato, P., van de Weg, W.E., Di Guardo, M., et al. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. Hortic. Res. **2**, 15060.

Nabwire, S., Suh, H.K., Kim, M.S., Baek, I., and Cho, B.K. (2021). Review: application of artificial intelligence in phenomics. Sensors **21**:4363.

NIST. (2015). https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf.

Niu, S., Liu, Y., Wang, J., and Song, H. (2020). A decade survey of transfer learning (2010–2020). IEEE Trans. Artif. Intell. **1**:151–166.

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J.M., Burgueño, J., Zhang, X., Singh, S., Vicente, F.S., Bonnett, D., Dreisigacker, S., et al. (2014). Genomic-enabled prediction with classification algorithms. Heredity **112**:616–626.

Ort, D.R., Merchant, S.S., Alric, J., Barkan, A., Blankenship, R.E., Bock, R., Croce, R., Hanson, M.R., Hibberd, J.M., Long, S.P., et al. (2015). Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proc. Natl. Acad. Sci. USA **112**:8529–8536.

Osnas, J.L.D. (2012). The extraordinary diversity of *Brassica oleracea*. https://botanistinthekitchen.blog/2012/11/05/the-extraordinary-diversityof-brassica-oleracea/.

Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F., Cornut, G., Costa, B.V., Ćwiek-Kupczyńska, H., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol. **227**:260–273.

Patten, B.C. (1998). Network orientors: steps toward a cosmography of ecosystems: orientors for directional development, self-organization, and autoevolution. In Eco Targets, Goal Functions, and Orientors, F. Muller and M. Leupelt, eds. (Berlin: Springer). p137-160.

Pazhamala, L.T., Kudapa, H., Weckwerth, W., Millar, A.H., and Varshney, R.K. (2021). Systems biology for crop improvement. Plant Genome **14**:e20098.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y., and Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 **2**:1595–1605.

Peng, B., Guan, K., Tang, J., Ainsworth, E.A., Asseng, S., Bernacchi, C.J., Cooper, M., Delucia, E.H., Elliott, J.W., Ewert, F., et al. (2020). Towards a multiscale crop modelling framework for climate change adaptation assessment. Native Plants **6**:338–348.

Peng, S., Khush, G.S., Virk, P., Tang, Q., and Zou, Y. (2008). Progress in ideotype breeding to increase rice yield potential. Field Crop. Res. **108**:32–38.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J **19**:3735–3747.

Piepho, H.-P. (2022). Prediction of and for new environments: what's your model? Mol. Plant **15**:581–582.

Pieruschka, R., and Schurr, U. (2019). Plant phenotyping: past, present, and future. Plant Phenomics **2019**:7507131.

Qian, Q. (2017). Smart super rice. Sci. China Life Sci. **60**:1460–1462.

Resende, R.T., Piepho, H.P., Rosa, G.J.M., Silva-Junior, O.B., e Silva, F.F., de Resende, M.D.V., and Grattapaglia, D. (2021). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor. Appl. Genet. **134**:95–112.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. **44**:217–220.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nat. Rev. Genet. **16**:85–97.

Rogers, A.R., Dunne, J.C., Romay, M.C., Bohn, M., Buckler, E.S., Ciampitti, I.A., Edwards, J., Ertl, D., Flint-Garcia, S., Gore, M.A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize. G3 **11**:jkaa050.

Runcie, D.E., Qu, J., Cheng, H., and Crawford, L. (2021). MegaLMM: mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biol. **22**:213.

Sadeghi-Tehran, P., Virlet, N., Ampe, E.M., Reyns, P., and Hawkesford, M.J. (2019). DeepCount: in-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks. Front. Plant Sci. **10**:1176.

Sandhu, K.S., Lozada, D.N., Zhang, Z., Pumphrey, M.O., and Carter, A.H. (2020). Deep learning for predicting complex traits in spring wheat breeding program. Front. Plant Sci. **11**:613325.

Sandhu, K., Patil, S.S., Pumphrey, M., and Carter, A. (2021). Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. Plant Genome **14**:e20119.

Sandhu, K.S., Patil, S.S., Aoun, M., and Carter, A.H. (2022a). Multi-trait multi-environment genomic prediction for end-use quality traits in winter wheat. Front. Genet. **13**:831020.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature **422**:297–302.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., and Melchinger, A.E. (2018). Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics **208**:1373–1385.

Schwab, K. (2017). The Fourth Industrial Revolution (Crown Business).

Selby, P., Abbeloos, R., Backlund, J.E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O.E., Birkett, C., Calaminos, V.C., Carceller, P., Cornut, G., et al. (2019). BrAPI – an application programming interface for plant breeding applications. Bioinformatics **35**:4147–4155.

Shahi, D., Guo, J., Pradhan, S., Khan, J., AVCI, M., Khan, N., McBreen, J., Bai, G., Reynolds, M., Foulkes, J., et al. (2022). Multi-trait genomic prediction using in-season physiological parameters increases prediction accuracy of complex traits in US wheat. BMC Genom. **23**:298.

Shalev-Shwartz, S., and Ben-David, S. (2014). Understanding Machine Learning from Theory to Algorithms (New York: Cambridge University press).

Sinha, P., Singh, V.K., Bohra, A., Kumar, A., Reif, J.C., and Varshney, R.K. (2021). Genomics and breeding innovations for enhancing genetic gain for climate resilience and nutrition traits. Theor. Appl. Genet. **134**:1829–1843.

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., and Deng, H.-W. (2020). A review of integrative imputation for multi-omics datasets. Front. Genet. **11**:570255.

South, P.F., Cavanagh, A.P., Liu, H.W., and Ort, D.R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. Science **363**:eaat9077.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L., and McCouch, S.R. (2015). Genomic selection and association mapping in rice (*Oryza Sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. **11**:e1004982.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**:1929–1958.

Staton, M., Cannon, E., Sanderson, L.-A., Wegrzyn, J., Anderson, T., Buehler, S., Cobo-Simón, I., Faaberg, K., Grau, E., Guignon, V., et al. (2021). Tripal, a community update after 10 years of supporting open source, standards-based genetic, genomic and breeding databases. Briefings Bioinf. **22**. bbab238-17.

Streich, J., Romero, J., Gazolla, J.G.F.M., Kainer, D., Cliff, A., Prates, E.T., brown, J.B., Khoury, S., Tuskan, G.A., Garvin, M., et al. (2020). Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? Curr. Opin. Biotechnol. **61**:217–225.

Sun, H., Jiao, W.-B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B., and Schneeberger, K. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat. Genet. **54**:342–348.

Tian, Z., Wang, J.-W., Li, J., and Han, B. (2021). Designing future crops: challenges and strategies for sustainable agriculture. Plant J. **105**:1165–1178.

Varshney, R.K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M.E. (2021). Designing future crops: genomics-assisted breeding comes of age. Trends Plant Sci. **26**:631–649.

Voss-Fels, K.P., Cooper, M., and Hayes, B.J. (2019). Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. **132**:669–686.

Wallace, J.G., Rodgers-Melnick, E., and Buckler, E.S. (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.

Wang, K., Shi, L., Liang, X., Zhao, P., Wang, W., Liu, J., Chang, Y., Hiei, Y., Yanagihara, C., Du, L., et al. (2022a). The gene TaWOX5 overcomes genotype dependency in wheat genetic transformation. Native Plants **8**:110–117.

Wang, N., Xia, X., Jiang, T., Li, L., Zhang, P., Niu, L., Cheng, H., Wang, K., and Lin, H. (2022b). In planta haploid induction by genome editing of DMP in the model legume *Medicago truncatula*. Plant Biotechnol. J. **20**:22–24.

Wang, R., Jiang, G., Feng, X., Nan, J., Zhang, X., Yuan, Q., and Lin, S. (2019a). Updating the genome of the elite rice variety Kongyu131 to expand its ecological adaptation region. Front. Plant Sci. **10**:288.

Wang, S., Wei, J., Li, R., Qu, H., Chater, J.M., Ma, R., Li, Y., Xie, W., and Jia, Z. (2019b). Identification of optimal prediction models using multiomic data for selecting hybrid rice. Heredity **123**:395–406.

Wang, X., Xuan, H., Evers, B., Shrestha, S., Pless, R., and Poland, J. (2019c). High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. GigaScience **8**:1–11.

Watanabe, K., Guo, W., Arai, K., Takanashi, H., Kajiya-Kanegae, H., Kobayashi, M., Yano, K., Tokunaga, T., Fujiwara, T., Tsutsumi, N., 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.

Watson, A., Ghosh, S., Williams, M.J., Cuddy, W.S., Simmonds, J., Rey, M.-D., Asyraf Md Hatta, M., Hinchliffe, A., Steed, A., Reynolds, D., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. Nat. Plants **4**:23–29.

Watt, M., Fiorani, F., Usadel, B., Rascher, U., Muller, O., and Schurr, U. (2020). Phenotyping: new windows into the plant for breeders. Annu. Rev. Plant Biol. **71**:689–712.

Westhues, M., Schrag, T.A., Heuer, C., Thaller, G., Utz, H.F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., et al. (2017). Omics-based hybrid prediction in maize. Theor. Appl. Genet. **130**:1927–1939.

Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., Liu, J., Wang, Q., Olsen, K.M., Han, B., et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding. Nat. Genet. **53**:243–253.

Westhues, C.C., Mahone, G.S., da Silva, S., Thorwarth, P., Schmidt, M., Richter, J.-C., Simianer, H., and Beissinger, T.M. (2021). Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. Front. Plant Sci. **12**:699589.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**:160018.

Wolpert, D.H. (1996). The lack of a priory distinction between learning algorithms. Neural Comput. **8**:1341–1390.

Wu, L., Han, L., Li, Q., Wang, G., Zhang, H., and Li, L. (2021). Using interactome big data to crack genetic mysteries and enhance future crop breeding. Mol. Plant **14**:77–94.

Xia, H., Zhang, X., Liu, Y., Bi, J., Ma, X., Zhang, A., Liu, H., Chen, L., Zhou, S., Gao, H., et al. (2022). Blue revolution for food security under carbon neutrality: a case from the water-saving and drought-resistance rice. Mol. Plant **15**:1401–1404.

Xiong, W., Reynolds, M., and Xu, Y. (2022). Climate change challenges plant breeding. Curr Opin Plant Biol. https://doi.org/10.1016/j.pbi.2022.102308.

Xu, H. (2020). Big data challenges in genomics. Handb. Stat. **43**:337–348.

Xu, Y. (2010). Molecular Plant Breeding (Wallingford: CAB International).

Xu, Y. (2015). Envirotyping and its applications in crop science. Sci. Agric. Sin. **48**:3354–3371.

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. Theor. Appl. Genet. **129**:653–673.

Xu, Y., and Crouch, J.H. (2008). Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. **48**:391–407.

Xu, Y., Lu, Y., Xie, C., Gao, S., Wan, J., and Prasanna, B.M. (2012). Whole-genome strategies for marker-assisted plant breeding. Mol. Breed. **29**:833–854.

Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., Prasanna, B.M., and Olsen, M.S. (2017). Enhancing genetic gain in the era of molecular breeding. J. Exp. Bot. **68**:2641–2666.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., Prasanna, B.M., Olsen, M.S., Wang, G., and Zhang, A. (2020a). Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun. **1**:100005.

Xu, Y., Yang, Q., Zheng, H., Xu, Y., Sang, Z., Guo, Z., Peng, H., Zhang, C., Lan, H., Wang, Y., et al. (2020b). Genotyping by target sequencing and its applications. Sci. Agric. Sin. **53**:2983–3004.

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C., et al. (2021a). Genomic selection: a breakthrough technology in rice breeding. Crop J. **9**:669–677.

Xu, Y., Zhao, Y., Wang, X., Ma, Y., Li, P., Yang, Z., Zhang, X., Xu, C., and Xu, S. (2021b). Incorporation of parental phenotypic data into multi-omic models improves prediction of yield-related traits in hybrid rice. Plant Biotechnol. J. **19**:261–272.

Xue, Y.B., Zhong, K., Han, B., Gui, J.F., Wang, T., Fu, X.D., He, Z.H., Chu, C.C., Tian, Z.X., Cheng, Z.K., et al. (2015). New chapter of designer breeding in China: update on strategic program of molecular module-based designer breeding systems. Bull. Chin. Acad. Sci. **30**:308–314. (in Chinese).

Yan, J., and Wang, X. (2022). Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology. Plant J. **111**:1527–1538.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X. (2021a). LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol. **22**:271.

Yan, S., Wang, S., Qiu, J., Li, M., Li, D., Xu, D., Li, D., and Liu, Q. (2021b). Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. Talanta **226**:122195.

Yan, W., Nilsen, K.T., and Beattie, A. (2022). Mega-environment analysis and breeding for specific adaptation. Crop Sci. (in press).

Yang, H.-W., Hsu, H.-C., Yang, C.-K., Tsai, M.-J., and Kuo, Y.-F. (2019a). Differentiating between morphologically similar species in genus *Cinnamomum* (Lauraceae) using deep convolutional neural networks. Comput. Electron. Agric. **162**:739–748.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019b). Federated machine learning. ACM Trans. Intell. Syst. Technol. **10**:1–19.

Yang, W., Guo, T., Luo, J., Zhang, R., Zhao, J., Warburton, M.L., Xiao, Y., and Yan, J. (2022). Target-oriented prioritization: targeted selection strategy by integrating organismal and molecular traits through predictive analytics in breeding. Genome Biol. **23**:80.

Yao, L., Zhang, Y., Liu, C., Liu, Y., Wang, Y., Liang, D., Liu, J., Sahoo, G., and Kelliher, T. (2018). OsMATL mutation induces haploid seed formation in indica rice. Nat. Plants **4**:530–533.

Yoosefzadeh-Najafabadi, M., Torabi, S., Tulpan, D., Rajcan, I., and Eskandari, M. (2021). Genome-wide association studies of soybean yield-related hyperspectral reflectance bands using machine learning-mediated data integration methods. Front. Plant Sci. **12**:777028.

Yu, H., and Li, J. (2022). Breeding future crops to feed the world through de novo domestication. Nat. Commun. **13**:1171.

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., Chen, M., Jing, Y., Kou, L., Li, X., et al. (2021a). A route to de novo domestication of wild allotetraploid rice. Cell **184**:1156–1170.e14.

Yu, S., Ali, J., Zhang, C., Li, Z., and Zhang, Q. (2020). Genomic breeding of green super rice varieties and their deployment in Asia and Africa. Theor. Appl. Genet. **133**:1427–1442.

Yu, S., Ali, J., Zhou, S., Ren, G., Xie, H., Xu, J., Yu, X., Zhou, F., Peng, S., Ma, L., et al. (2021b). From Green Super Rice to green agriculture: reaping the promise of functional genomics research. Mol. Plant **15**:9–26.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., and Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere **10**:373.

Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., Zhu, G., Xiong, X., Shang, Y., Li, C., et al. (2021a). Genome design of hybrid potato. Cell **184**:3873–3883.e12.

Zhang, X., Qian, Q., Zhang, J., Dreng, X.W., Wan, J., and Xu, Y. (2021b). Transforming and upgrading off-season breeding in Hainan through molecular plant breeding. Sci. Agric. Sin. **54**:3789–3804.

Zhao, X., Meng, Z., Wang, Y., Chen, W., Sun, C., Cui, B., Cui, J., Yu, M., Zeng, Z., Guo, S., et al. (2017). Pollen magnetofection for genetic modification with magnetic nanoparticles as gene carriers. Nat. Plants **3**:956–964.

Zheng, Z., Wang, H.B., Chen, G.D., Yan, G.J., and Liu, C.J. (2013). A procedure allowing up to eight generations of wheat and nine generations of barley per annum. Euphytica **191**:311–316.