




















Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates

Jacob D. Washburn ^{1,*}, José Ignacio Varela^{2,3,†}, Alencar Xavier ^{3,4,†}, Qiuyue Chen ⁵, David Ertl⁶, Joseph L. Gage ⁵, James B. Holland ^{5,7}, Dayane Cristina Lima², Maria Cinta Romay ⁸, Marco Lopez-Cruz ⁹, Gustavo de los Campos ⁹, Wesley Barber³, Cristiano Zimmer ³, Ignacio Trucillo Silva³, Fabiani Rocha³, Renaud Rincint ¹⁰, Baber Ali¹⁰, Haixiao Hu ¹¹, Daniel E. Runcie ¹¹, Kirill Gusev¹², Andrei Slabodkin¹², Phillip Bax¹², Julie Aubert¹³, Hugo Gangloff¹³, Tristan Mary-Huard^{10,13}, Theodore Vanrenterghem¹³, Carlos Quesada-Traver ¹⁴, Steven Yates¹⁴, Daniel Ariza-Suárez¹⁴, Argeo Ulrich^{15,16}, Michele Wyler ¹⁷, Daniel R. Kick ¹, Emily S. Bellis¹⁸, Jason L. Causey¹⁸, Emilio Soriano Chavez¹⁸, Yixing Wang¹⁸, Ved Piyush¹⁹, Gayara D. Fernando¹⁹, Robert K. Hu²⁰, Rachit Kumar^{20,21}, Annan J. Timon²⁰, Rasika Venkatesh²⁰, Kenia Segura Abá^{22,23}, Huan Chen²³, Thilanka Ranaweera^{22,24}, Shin-Han Shiu ^{22,24,25}, Peiran Wang^{26,27}, Max J. Gordon^{26,27}, B Kirtley Amos^{26,27}, Sebastiano Busato^{26,27}, Daniel Perondi^{26,27}, Abhishek Gogna ²⁸, Dennis Psaroudakis²⁹, Chun-Peng James Chen³⁰, Hawlader A. Al-Mamun³¹, Monica F. Danilevicz³¹, Shriprabha R. Upadhyaya³¹, David Edwards ³¹, Natalia de Leon ^{2,*}

¹USDA-ARS, MWA-PGRU, 302-A Curtis Hall, University of Missouri, Columbia, MO 65211, USA

²Department of Plant and Agroecosystem Sciences, University of Wisconsin—Madison, 1575 Linden Drive, Madison, WI 53706, USA

³Corteva Agrisciences, 8305 NW 62nd Ave, Johnston, IA 50131, USA

⁴Department of Agronomy, Purdue University, 915 Mitch Daniels Blvd, West Lafayette, IN 47907, USA

⁵Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA

⁶Iowa Corn Promotion Board, Johnston, IA 50131, USA

⁷USDA-ARS, Plant Science Research Unit, Raleigh, NC 27695, USA

⁸Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

⁹Departments of Epidemiology and Biostatistics and Statistics and Probability, and Institute for Quantitative Health Science and Engineering, Michigan State University, 775 Woodlot Dr, East Lansing, MI 48823, USA

¹⁰Université Paris—Saclay, INRAE, CNRS, AgroParisTech, GQE—Le Moulon, 91190 Gif-sur-Yvette, France

¹¹Department of Plant Sciences, University of California Davis, One Shield Drive, Davis, CA 95616, USA

¹²Smart Agri Labs, 2055 Limestone Rd STE 200-C, Wilmington, DE 19808, USA

¹³Université Paris—Saclay, AgroParisTech, INRAE, UMR MIA Paris—Saclay, 91120 Palaiseau, France

¹⁴Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitätsstrasse 2, CH-8092 Zurich, Switzerland

¹⁵Puregene AG, Etzmatt 273, CH-4314 Zeiningen, Switzerland

¹⁶Institute of Agricultural Sciences, ETH Zurich, Universitätsstrasse 2, CH-8092 Zürich, Switzerland

¹⁷MWSchmid GmbH, Hauptstrasse 34, CH-8750 Glarus, Switzerland

¹⁸Department of Computer Science, Arkansas State University, 2105 E. Aggie Rd, Jonesboro, AR 72401, USA

¹⁹Department of Statistics, University of Nebraska—Lincoln, 340 Hardin Hall North Wing, Lincoln, NE 68583, USA

²⁰Genomics and Computational Biology, Perelman School of Medicine at the University of Pennsylvania, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, USA

²¹Medical Scientist Training Program, Perelman School of Medicine at the University of Pennsylvania, University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19104, USA

²²DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA

²³Genetics and Genome Sciences Graduate Program, Michigan State University, East Lansing, MI 48824, USA

²⁴Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

²⁵Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

²⁶NC Plant Science Initiative, North Carolina State University, 840 Oval Drive, Raleigh, NC 27606, USA

²⁷Department of Electrical and Computer Engineering, North Carolina State University, 890 Oval Dr, Raleigh, NC 27606, USA

²⁸Department of Breeding Research, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Corrensstraße 3, Gatersleben 6466, Germany

²⁹Department of Molecular Genetics, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Corrensstraße 3, Gatersleben 6466, Germany

³⁰School of Animal Sciences, Virginia Tech, Blacksburg, VA 24061, USA

³¹School of Biological Sciences and Centre of Applied Bioinformatics, University of Western Australia, Perth, WA 6009, Australia

*Corresponding author: USDA-ARS-MWA-PGRU, 302-A Curtis Hall, University of Missouri, Columbia, MO 65211, USA. Email: jacob.washburn@usda.gov; *Corresponding author: Department of Plant and Agroecosystem Sciences, University of Wisconsin—Madison, 1575 Linden Drive, Madison, WI 53706, USA. Email: ndeleongatti@wisc.edu

[†]These authors contributed equally to this work.

Predicting phenotypes from a combination of genetic and environmental factors is a grand challenge of modern biology. Slight improvements in this area have the potential to save lives, improve food and fuel security, permit better care of the planet, and create other positive outcomes. In 2022 and 2023, the first open-to-the-public Genomes to Fields initiative Genotype by Environment prediction

competition was held using a large dataset including genomic variation, phenotype and weather measurements, and field management notes gathered by the project over 9 years. The competition attracted registrants from around the world with representation from academic, government, industry, and nonprofit institutions as well as unaffiliated. These participants came from diverse disciplines, including plant science, animal science, breeding, statistics, computational biology, and others. Some participants had no formal genetics or plant-related training, and some were just beginning their graduate education. The teams applied varied methods and strategies, providing a wealth of modeling knowledge based on a common dataset. The winner's strategy involved 2 models combining machine learning and traditional breeding tools: 1 model emphasized environment using features extracted by random forest, ridge regression, and least squares, and 1 focused on genetics. Other high-performing teams' methods included quantitative genetics, machine learning/deep learning, mechanistic models, and model ensembles. The dataset factors used, such as genetics, weather, and management data, were also diverse, demonstrating that no single model or strategy is far superior to all others within the context of this competition.

Keywords: genotype by environment; prediction; competition; maize; yield; phenotype

Introduction

Phenotype prediction is a grand challenge of 21 century biology (National Research Council (US) 2010; Azodi et al. 2019; Martinez 2023; US National Science Foundation 2023). The ultimate goal of plant and animal breeding is to develop better phenotypes. Historically, breeding has been based on observation of phenotypes, recombining elite genetics, and then selecting for superior phenotypes. With the advent of molecular genetic marker technologies and DNA sequencing, phenotypic selection has been augmented with genomic prediction (GP) and selection. In an agricultural context, GP and genomic selection (GS) have fundamentally altered plant and animal breeding by reducing generation interval, predicting traits that are too difficult and/or expensive to measure at population scale, and in the case of animals, improving the selection of sex-limited traits (Meuwissen et al. 2001; Heffner et al. 2009; Lorenz et al. 2011; Desta and Ortiz 2014; Bhat et al. 2016; Crossa et al. 2017; Wiggans et al. 2017; Washburn et al. 2020; Budhlakoti et al. 2022; Johnsson 2023). Similarly, prediction models based on environmental and/or agronomic management factors have been critical tools for crop-risk assessment, fertilizer and irrigation prescription, sustainability forecasting, climate change modeling, and other basic and applied research and decision-making (Hammer et al. 2002, 2019; Jones et al. 2003; Keating et al. 2003; Archontoulis et al. 2014; Di Paola et al. 2016; Schauburger et al. 2017; Challinor et al. 2018).

The terminology and importance placed on different factors can vary across applications. For example, plant and animal breeding tend to focus on the influence of genetic (G) factors on phenotypes, while sometimes including environmental (E) factors, and genotype-by-environment (G×E) interactions in their models. These fields take a plant-centric viewpoint where agronomic management (M) is included within E. Disciplines like agronomy and plant physiology, on the other hand, often focus on M factors as distinct from E, taking a more farmer-centric approach where E is uncontrollable, and M includes all factors that can be manipulated by the researcher or producer. G may even be included as part of M in this framework, as the producer decides which cultivars to plant, in the same way that they determine fertilizer and irrigation rates.

Historically, applied agronomic prediction methods have focused on G, E, and/or M factors in relative isolation, but the past decade has seen a renewed interest in, and recognition of the need for, models that incorporate G, E, and M factors within a unified framework (Jarquín et al. 2014; Technow et al. 2015; Messina et al. 2018, 2023; Li et al. 2021; Washburn et al. 2021; Cooper et al. 2022; Guo and Li 2023; Kick et al. 2023; Kick and Washburn 2023; Lopez-Cruz et al. 2023). These approaches are sometimes referred to as G×E or G×E×M depending on the context, but in practice,

they rely on diverse combinations of factors modeled in interacting and noninteracting ways.

In breeding, GP and GS approaches are well-suited for complex traits such as yield, which tend to have small single genetic variant (and per gene) effect sizes. They rely on genetic variation across the genome, as opposed to marker-assisted selection approaches that use only a few selected loci known to be predictive of important variation for the trait of interest (Haley and Visscher 1998; Meuwissen et al. 2001; Heffner et al. 2009). Although many methods have been developed over the years to improve GP/GS accuracy, common linear and nonlinear GP models, including Genomic Best Linear Unbiased Prediction (GBLUP), Bayesian algorithms, Ridge Regression, Artificial Neural Networks, and Decision Tree-based algorithms, have often been shown to perform similarly across different species and trait combinations in cases where only G data are used, but differences can be large in some cases (Azodi et al. 2019; Charmet et al. 2020; Montesinos-López et al. 2021). Many genomic approaches have been developed to predict nonadditive genetic effects, including modeling additive and dominant effects of individual markers, using multiple kernels (e.g. additive and dominance kernels; Vitezica et al. 2013), and nonlinear kernel regression methods (e.g. Reproducing Kernel Hilbert Spaces; Morota and Gianola 2014). However, linear models often perform similarly to, or even outperform, nonlinear algorithms, particularly when the trait has a predominantly additive genetic basis (Azodi et al. 2019; Montesinos-López et al. 2021).

When environmental variation is included in the model, linear, nonlinear, and combined approaches that explicitly model environmental factors either statistically or mechanistically have been shown to improve predictive accuracy, particularly under certain stressful environments (Jarquín et al. 2014; Technow et al. 2015; Ly et al. 2017; Millet et al. 2019; Li et al. 2021; Washburn et al. 2021; Diepenbrock et al. 2022). Due to the complexity of experimental design in plant breeding trials, other practical strategies for analyzing multi-environment data have been developed, such as fitting linear mixed models with a 2-stage approach for accelerating the speed of computation (Möhring and Piepho 2009; Rogers et al. 2021). The inclusion of dominance effects for hybrids and G×E effects for environment-specific GP are also likely to increase accuracy (Rogers et al. 2021; Rogers and Holland 2022). Another way to analyze multi-environmental trials is to group environments into what are termed mega environments (Lin et al. 2021).

Significant efforts have been made in the area of machine learning models for agriculture. Random forest (RF) models have been used to predict and analyze complex traits in plants by leveraging the collective intelligence of multiple decision trees (Azodi et al. 2019). RF excels at handling high-dimensional data and capturing nonlinear relationships between predictors and responses.

These attributes make it more suitable for capturing nonadditive genetic effects and potentially environmental variation. Additionally, RF can provide variable importance measures, enabling breeders to identify key genetic markers and prioritize traits for selection. This versatility and robustness have made RF a valuable go-to tool in prediction (Montesinos López et al. 2022). RF models have outperformed other methods for some traits and in some contexts of GP (González-Recio and Forni 2011; Charmet et al. 2020; Montesinos López et al. 2022).

Boosting is another method that is highly effective and is becoming widely used. Extreme gradient boosting is a scalable tree boosting system that has been recognized for its computational speed and accuracy across a range of prediction problems (Chen and Guestrin 2016). Gradient boosting of decision trees, for example, XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al. 2017), and related RF methods have been observed to significantly outperform deep learning models in several plant and nonplant tabular datasets and they typically require fewer computational resources (Danilevicz et al. 2021; Borisov et al. 2022; Gill et al. 2022; Grinsztajn et al. 2022; Shwartz-Ziv and Armon 2022).

Predicting phenotypes often requires combining multiple heterogeneous sources of information (Xu et al. 2019). This can complicate the selection of a meaningful subset of features to use as predictors. Deep learning models are highly effective in handling large and diverse input data and modeling complex nonlinear relationships, surpassing traditional modeling approaches (Khaki and Wang 2019). Additionally, they can incorporate specialized architectures such as long-short-term memory, which is particularly suitable for time series data due to its capacity to capture and retain long-term dependencies (Malhotra et al. 2015; Shook et al. 2021). Deep learning models have shown mixed results when applied to GP scenarios (Montesinos-López et al. 2021; Washburn et al. 2021; Kick et al. 2023). Most successful applications of deep learning involve enormous datasets of a scale beyond those typically used in plant breeding and other agricultural scenarios, which has potentially limited the true potential of these methods in agricultural phenotype prediction. The computational requirements of these methods are also potentially limiting even with relatively small datasets, but the computer gaming industry has resulted in wider access to inexpensive graphical processing units, making these methods arguably less expensive than some traditional GBLUP approaches that require prohibitively large amounts of memory. Other approaches have involved a combination of deterministic models, expert knowledge, and deep learning to model environmental stress in agriculture (Cvejoski et al. 2021).

Another important class of methods is ensembles of different model types. Ensembles allow the combination of other types of methods, including any of those discussed above, into a single prediction and can often outperform each method on its own (Shahhosseini et al. 2020; Kick and Washburn 2023). Ideally, ensembling allows for reduced error (Zhou 2015) and greater robustness by pooling the predictions of a diverse set of models. Ensemble learning has been used for purposes ranging from optimizing genetic transfer to genetic prediction, but substantially, more research has focused on developing and comparing single model performance than ensemble approaches (Azodi et al. 2019; Hesami et al. 2020; Liang et al. 2021).

One significant challenge to improving phenotype prediction methods is the lack of publicly available datasets containing G, E, and M factors with which to experiment and develop new prediction approaches. In 2013, collaborators from universities, government, farmer associations, and industry recognized the need

for this type of data and formed the Genomes to Fields (G2F) Initiative Maize Genomes by Environment (GxE) project centered around maize yield trial data. To date, this project has evaluated over 180,000 unique plots, 5,000 maize hybrids, and 280 environments (McFarland et al. 2020; Lima et al. 2023a, 2023b). Phenotypic, genetic, environmental, and management data have been collected following standard protocols and uploaded annually to a joint repository for public use. The individual datasets have resulted in many publications and advancements in our knowledge and understanding of maize genetics, phenotypes, and environmental responses (DeChant et al. 2017; Gage et al. 2017, 2019; Wiesner-Hanks et al. 2018, 2019; Bai et al. 2019; Stewart et al. 2019; Wu et al. 2019; Falcon et al. 2020; Morales et al. 2020; Sekhon et al. 2020; Rogers et al. 2021; Lopez-Cruz et al. 2023). The datasets have also been used individually for multiple studies on phenotype prediction (Anderson et al. 2019; Anche et al. 2020; Jarquin et al. 2021; Washburn et al. 2021; Westhues et al. 2021, 2022; Rogers and Holland 2022; Kick et al. 2023; Kick and Washburn 2023; Winn et al. 2023).

While the G2F dataset has been shared with the public throughout the project, and continues to be updated as new data is collected and processed, barriers to using the data have persisted (Lopez-Cruz et al. 2023). These barriers included the vast size and complexity of the dataset, the many different data types collected, each of which requires some unique domain knowledge to interpret, and the fact that many potential users were simply unaware of the data. Additionally, to make the dataset useful for a wide range of studies, curation has been kept to a minimum to allow each researcher to determine what data are most valid for their specific use case.

One of the original goals of the G2F GxE project had been to use the data for the development of phenotype prediction methods. From 2022 November 15 to 2023 January 15, the G2F GxE hosted its first ever yield prediction competition with the objectives of expanding the number of people and domains working with the dataset and stimulating the development of new and innovative phenotype prediction models and strategies. Competition challenges are very common in some domains of science (e.g. computer science, data science, etc.), but they are relatively uncommon in plant science, breeding, genetics, and related disciplines. For the G2F community, this was a completely new endeavor. Planning and advertising for the competition began more than a year in advance, along with an extensive effort to format and curate the existing G2F data to date into a form that would be more accessible to competition participants and future researchers. The goal for competition participants was to use the G2F GxE data from 2014 to 2021 (termed the training set) to predict the grain yield results from the G2F GxE 2022 season (termed the testing set) which had not yet been publicly released. This represented one of the most difficult prediction scenarios in crop agriculture because it required prediction on many new genotypes (hybrids not included in the training set) in new environments (a new year not included in the training set). Of the hybrids for prediction in 2022, only 8% had been previously tested in the G2F GxE project. Many of the field locations had been used previously, but no yield performance data at any of the locations for 2022 was provided to the participants.

This manuscript describes the results of the competition, including the different modeling strategies employed and their effectiveness. To encourage industry participation, teams were allowed to keep their methods confidential (see Materials and methods for details). Teams were also allowed to publish their models on their own rather than being included in this joint

manuscript. Several teams elected to do so and their manuscripts are in various stages of preparation, submission, and publication along with nonparticipants using the data independently (Fernandes et al. 2024; Ge et al. 2024; Khalilzadeh et al. 2024; Lopez-Cruz et al. 2024). This manuscript includes the results from all eligible participant teams as well as detailed methods, results, and code for 17 of the participant teams, including a majority of the top 10 and others distributed across the range of scores, who opted to participate. Particular emphasis is given to the winning team's approach, methods, and results.

Materials and methods

Competition datasets

A substantial curation, quality control (QC), formatting, and standardization effort on the G2F data were carried out to lower the barrier to entry and facilitate the participation of individuals from as many different disciplines and backgrounds as possible. Compilation of the phenotypic dataset began by combining the “clean” versions of the raw trait data files posted as DOIs for public distribution. The “clean” trait data files had undergone automated filtering to remove extreme outliers, as described in the DOI readme files and dataset publications (Lima et al. 2023a, 2023b, 2023c). Trait data only from the “main” G2F experiments were included; data from several smaller “side” experiments were removed. Subexperiment and field block information were recovered for missing information whenever possible. Trait and meta-information column names were harmonized among years. Hybrid names were harmonized between years of the training data and between training and testing sets. Duplicate records were dropped.

Metadata files were curated in a similar way as the phenotypic data; files were downloaded from DOIs and compiled. Environment names were harmonized among the years, and between metadata and phenotypic data to ensure consistency. Only those environments present in the phenotypic data were kept in the metadata. Any specific information related to a particular environment was added to the “Comments” column. Columns for “City,” “Farm,” and “Previous Crop” were also harmonized, double-checked, and updated, when possible. Soil data collection only began in 2015. For all years after that, files were obtained from the DOI and compiled. As with the metadata, environment names were standardized across all years and between phenotypic data, and only environments present in the phenotypic data were included in the soil data file. For some environments, data were unavailable in the DOI, as the results were returned after the DOI release. In such cases, the data were retrieved from the laboratory responsible for the analysis and included in the competition DOI release (Lima et al. 2023c).

For the genotype data, the maize practical haplotype graph (PHG) was used for variant calling (Bradbury et al. 2022). The PHG aligns sequencing reads with genome assembly sequences to impute genotypes based on stored haplotypes. The Maize 2.1 PHG haplotypes came from 86 genome assemblies which were aligned to the B73 v5 assembly (Hufford et al. 2021) with anchor-wave (Song et al. 2022). The assemblies came from MaizeGDB and other sources (Yang et al. 2019; Bornowski et al. 2021; Woodhouse et al. 2021). The B73 genome was divided into nodes (both genic and intergenic) known as reference ranges, using the annotations from Zm-B73-REFERENCE-NAM-5.0_Zm00001eb.1.gff3. The ends of reference ranges were selected as regions with 10 or more conserved base pairs in 23 or more of the 25 NAM genome assemblies (Hufford et al. 2021). Genic haplotypes having 0.0001

or lower divergence and intergenic haplotypes having 0.001 or lower divergence were collapsed into consensus haplotypes.

Due to changing technologies over the years, the inbred sequence data (which is combined to create the hybrid genotypes) collected for the G2F project have come from different methods in different years of the project. The 2014–2017 genotyping was done using a GBS (genotyping by sequencing) protocol (Elshire et al. 2011) with the ApeKI restriction enzyme. The 2018–2019 genotyping was performed with ~5× coverage skim sequence. The 2020–2021 genotyping used exome capture in combination with GBS controls using ApeKI. The 2022–2023 genotyping used GBS with PstI-MspI. To create the genotype calls from the G2F data, inbred reads were aligned to the PHG to identify haplotypes matching each read. The haplotype path through the graph was then imputed and used to identify variants. Only positions contained within the 600k SNP genotyping array (Unterseer et al. 2014) were considered. CrossMap was used to uplift array positions to v5 coordinates, and positions were dropped if they could not be uplifted, were missing in 21 or more assemblies, or were monomorphic in all assemblies (Zhao et al. 2014). The final set contained 437,214 variant positions. Hybrid names in the genotypic dataset were harmonized with the hybrid names in training and testing trait datasets.

Weather data were downloaded from the NASA Power website (<https://power.larc.nasa.gov/>) for the locations and years in the training and testing sets. An estimate was made for locations where the exact GPS field coordinates are unknown. Competitors could also use other sources of weather data, including the weather data available in the DOI of each year of the project; each environment is equipped with a weather station (WatchDog 2700 Weather Station) that records various weather parameters every 30 min during the growing season, including air temperature, humidity, solar radiation, rainfall, wind speed and direction, soil temperature, and soil moisture. Environmental covariate (EC) data were also given to the participants. Details of how these ECs were generated are described in Lopez-Cruz et al. (2023). It is important to note that these covariates were defined at the year-location level; therefore, these did not vary within year-location. Another important detail about the training and testing environments is that most of the locations were identical across years. The exact fields used were often different to enable crop rotation and other management practices, but evaluations at any given location were typically near those of previous years. Most locations and years had GPS coordinate data that were provided to the participants as metadata.

The final curated dataset used in the competition, including the observed phenotypic (test set) values that were not released until after the competition, is publicly available for further use (Lima et al. 2023c). The released dataset also contains an extensive readme file with details about each component of the dataset exactly as they were presented to the competition participants. The training and testing set data and a short numerical summary of the contents of each are summarized in Table 1.

Competition procedures and methods

The competition hosted a website for advertising, information, and competition rules and utilized the EvalAI platform (Yadav et al. 2019) for submissions, evaluations, and leaderboard hosting (see Supplementary File 1 which contains a preserved version of all competition website information). EvalAI is an open-source platform for evaluating and comparing machine learning and artificial intelligence outputs and algorithms. The platform allows the host to control the number of challenge phases, dataset splits,

Table 1. Data files made available for the competition.

Set, years	File description	Unique environments	Unique hybrids	Variant sites
Training, 2014–2021	Trait, hybrid, and experimental design details	217	4,683	N/A
Training, 2014–2021	Individual trial metadata including locations, management, treatments, etc.	217	N/A	N/A
Training, 2014–2021	Soil testing data from individual trial locations	141	N/A	N/A
Training, 2014–2021	Weather data for trial locations as retrieved from NASA power	212	N/A	N/A
Training, 2014–2021	ECs calculated using the APSIM crop model	165	N/A	N/A
Training and testing, all years	Genotypic data for all public hybrids in the competition	N/A	4,928	437,214
Testing, 2022	Submission template file with testing set hybrid and environment names	26	548	N/A
Testing, 2022	Individual trial metadata including locations, management, treatments, etc.	26	N/A	N/A
Testing, 2022	Soil testing data from individual trial locations	21	N/A	N/A
Testing, 2022	Weather data for trial locations as retrieved from NASA power	26	N/A	N/A
Testing, 2022	ECs calculated using the APSIM crop model	24	N/A	N/A

and leaderboard visibility by cloning a git repository that contains the configuration files in the YAML language. The competition rules were described in detail in the website materials and the EvalAI challenge webpage. Individuals and teams from any institution were encouraged to participate but could only be part of 1 team in the competition. Any model or strategy was permissible as long as it relied only on data provided by the competition or external data that were publicly available by 2022 February 1. This date was chosen to prevent competitors from using private data or data collected during the 2022 growing season, which is the season they were trying to predict. The submitted predictions were to be of absolute grain yield for each hybrid for each test environment. These were reported in megagrams (metric tons) per hectare (Mg/ha) with the standard 15.5% moisture adjustment used for maize in the United States. The evaluation metric used was the average of the root MSE (RMSE) calculated for each environment. In other words, for each submission, the RMSE for each environment was calculated individually, and then these RMSE values were averaged for a final score. The winner was the team with the lowest average RMSE value. To be eligible to win the competition and receive the cash prize, participants were required to commit to publishing their model code and results after the competition (on their own or as part of this combined manuscript). To allow greater participation from industry groups, an option was provided for participation in the competition without publication of methods, but these groups would be ineligible for the prize and winning title.

The competition began on 2022 November 15 at which time the training data, testing data (with the exception of yield), and submission template (Table 1) were provided to all teams. Each team was allowed to make 5 total submissions through the EvalAI system by the close of the competition on 2023 January 15. Results (mean RMSE) were evaluated automatically using a Python script hosted in the EvalAI remote servers and displayed on the competition leaderboard within seconds of submission. The calculation of RMSE was performed within each environment and then averaged across all environments. The EvalAI leaderboard provided participants with feedback on their submissions that could be used to improve their model and also allowed for greater interaction between participants. At the end of the competition, the top 3 teams on the leaderboard were contacted and required to provide their code to the organizers for a complete evaluation. This process ensured that the winning team was following competition rules as outlined and their results could be

re-created in the hands of the organizers. There was 1 team (DataJanitors) made up of a few of the organizers of the competition; they were allowed to make submissions “for fun” as long as they followed all the normal competition rules (not using the truth values), but they would not and were not considered part of the competition or eligible to win. However, their methods and results are included here.

Individual team methods

The prediction methods used by different teams in the competition were extremely diverse. Methods ranged from traditional GBLUP and linear mixed model approaches, to deterministic models, to machine learning methods, RF models, ridge regression, gradient boosted decision trees, and various deep learning neural network approaches (Breiman 2001; Bradbury et al. 2007; Pedregosa et al. 2011; Jarquín et al. 2014; Pérez and de los Campos 2014; Chollet 2015; Abadi et al. 2016; Chen and Guestrin 2016; Butler et al. 2017; Ke et al. 2017; Wright and Ziegler 2017; Paszke et al. 2019). Many of the methods were also applied jointly using ensemble approaches. Most methods were implemented using packages in Python and R (Van Rossum and Drake 2009; R Core Team 2021). Some teams used feature selection methods, GWAS, and/or external datasets with no direct relationship to the G2F data in attempts to improve their model's predictive accuracy. The methods and approaches used by each team are described in detail in [Supplementary File 2](#) and links are provided there to the code developed by each participating team.

Winning team methods

The most accurate prediction model in the competition was submitted by the Corteva Latin America Corn (CLAC) team. It consisted of averaging 2 model outputs, namely A and B. Model A was a univariate linear mixed-effects model, where the fixed effects included location metadata: location (state, station), irrigation, treatment, and previous crop. No year or year-location effects were accounted for. The random effect consisted of a polygenic genetic term, with a relationship matrix calculated as an arc-cosine kernel. Predictions from model A were generated from the fixed effects and genomic values estimated for the 2022 data.

Model B consisted of a location-specific model using environmental variables and the location metadata to predict the environmental means, while relying on an index from a multivariate GBLUP to predict the genetic component. The environment means

model fitted the mean yield of environments as a function of environmental variables and the location metadata, fitted in 2 steps: first, it estimated the biased composite predictor as the average of 3 submodels: (1) an RF of the environmental factors, (2) a ridge regression of environmental factors, (3) a least squares of the location metadata. Second, it computed the unbiased estimator, using a linear regression to remove the shrinkage from the composite prediction. Model B's predicted genomic values were inferred from a selection index. The predictions started from fitting an unstructured model of the observed environments as

$$\begin{aligned}y^* &= g + e \\g &= \Sigma_g \otimes K \\e &= \Sigma_e \otimes I\end{aligned}$$

where y^* was standardized and spatially adjusted phenotypic values; g and e were the genetic effect and residuals of each corresponding phenotype, respectively. The genetic covariance matrix Σ_g contained the variances of each environment in the diagonal and covariances between pair of environments in the off diagonal. The residual covariance Σ_e was a diagonal matrix containing the residual variance for each environment. K was the genomic relationship matrix, where the pairwise relationship among individuals was calculated from an arc-cosine function using genomic information (Cuevas et al. 2019). The output of the model consisted of predictions of every individual in every environment. The predicted genetic merit for the k th 2022 location (u_k) was estimated as a linear combination of the genomic values of the observed locations as:

$$u_k = \sigma_k \sum_{i=1}^I g_i w_{i,k}$$

where the scalar $w_{i,k}$ corresponded to the weight of the i th location to predict the k th location of 2022, and σ_k was the predicted standard deviation for the k th location. The genotypic standard deviation of 2022 locations was predicted from the phenotypic standard deviation of location as a function of ECs using RF ($\sigma = \text{RF}(\mathbf{W}) + e$). The weights ($w_{i,k}$) were based on the deterministic accuracy of i th predicting k th location and the geographical location, such that locations in the same state and station would have higher weights than location further away, and locations with

individuals more related to those in the k th location will also have higher weights.

All computations were done in R. Linear mixed models were fitted using the R package bWGR 2.1 (Xavier 2019; Xavier et al. 2020; Xavier and Habier 2022). The univariate model was described by Xavier (2019) and the multivariate model by Xavier and Habier (2022). The RF used the R package ranger (Wright and Ziegler 2017). Full code for running the model is provided with this paper at: https://github.com/alenvxav/Lectures/tree/master/MGC_2023.

Results and discussion

Competition participation

Overall, the competition had 241 registrants comprising 128 teams from at least 19 countries and every inhabited continent (Fig. 1a). Registrants came from academic institutions, government, nonprofits, and private companies (including major seed companies, small startups, biotechnology companies, and international tech giants, Fig. 1b). The mean number of members of a team was 1.9 with most teams consisting of only one individual and the largest team consisting of 10. Other demographic information on the participants was not recorded. By the end of the competition, 30 teams (excluding a few test teams and erroneous submission) successfully submitted a model on the leaderboard. That represents a 77% drop out rate, which was congruent with expectations given the significant effort and time required to develop a successful model with a large and heterogeneous dataset of this kind. Interestingly, while 62% of the original registered teams contained only one member, 70% of the teams with successful submissions had 2 or more members and the mean number of members in these final teams was 2.9.

Model accuracies

One important difference between this competition and other model development studies is the fact that participants did not have access to the observed yield values in the test set. The best practice is for researchers to have a "validation set" in addition to training and testing sets, and iteratively improve their models on that set, saving the testing set for final testing and evaluation only. Ideally, the number of times a model is challenged with the testing set (and potentially changed by the researcher to perform better) should be minimized. However, this is difficult to

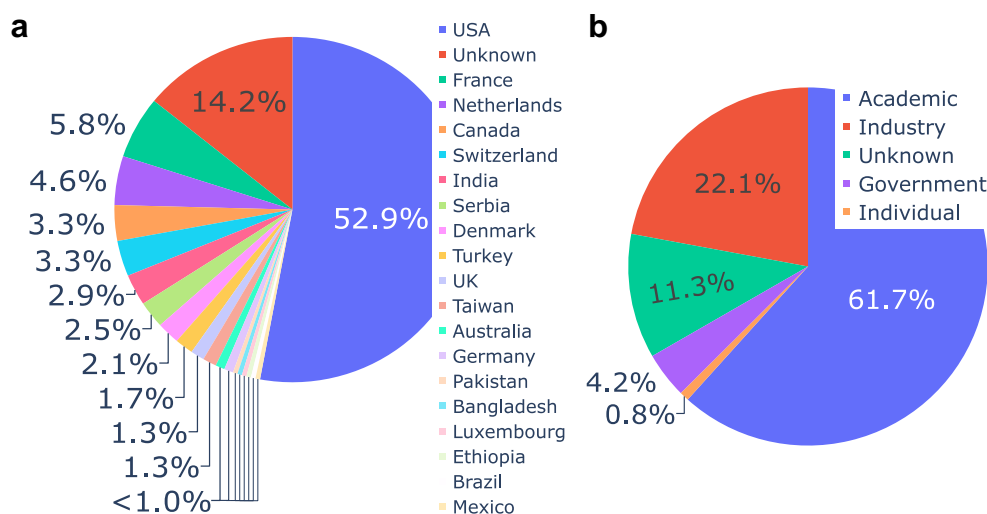


Fig. 1. Competition participation by: a) country and b) institution type.

implement for even the most well-meaning researchers due to small sample sizes, concerns about representation in sets, and the fact that if one's model performs very poorly on the testing set, it may be unpublishable or indicative of improvement areas. In this competition, participants had up to 5 chances to challenge their model against the testing set; 4 of these could be used as feedback to improve the model for the fifth attempt. The motivation for this is described in the Materials and methods section. While this allowed for some improvement based on testing set feedback, the number of possible iterations was much lower than what one could do with testing set access, and the feedback given was minimal (a single accuracy score).

Of the teams that submitted models, 67% made all 5 submissions, 10% made 4, 7% made 3, 3% made 2, and 13% made only one. There was a significant ($P = 0.024$) Spearman correlation of -0.412 between the final rankings of teams and the number of submissions they made, indicating that teams with more submissions did better on average than teams with fewer (the small sample size of 30 teams should be considered in these interpretations). Even so, it did not appear that having 5 submissions played a major role in model improvements. Most submissions were made near the end of the competition (see [Supplementary Fig. 1](#)) with 71% of submissions in the final week and almost 60% in the last 3 days. When submissions from each team were compared with their previous submission (e.g. submission 5 vs submission 4, 4 vs 3, etc.), 52% of submissions performed better than the previous submission and 48% performed worse. When only the first and final per team submissions were considered, 46% of final submissions resulted in improvement, while 54% resulted in lower accuracy. The reasons behind this are not clear, but it may have been that more committed teams simply made more submission attempts. The winning team was in first place for most of the competition with only a few days when the second-place team was in the lead.

The density distribution of the final model accuracy scores from the competition is shown in [Fig. 2](#). Only the top 10 models are highlighted in the figure with an "x". The top 10 best scoring models are relatively close together: within 0.215 Mg/ha. As a percentage of the average yield of the testing set, this difference in model errors is 2.13%, indicating a small improvement from the

10th place model to the 1st place, but one that could still be considered useful for many applications. Looking across the competition (with the exclusion of 1 outlier model), the 1st ranked team's score represents a 9.42% improvement.

Different accuracy metrics

While many measures of model accuracy could have been used, the average RMSE metric was chosen for determining the competition winner due to its simplicity and common use in modeling, machine learning, and deep learning (from which the organizers hoped to attract participants). The best evaluation metric in any study depends on the goals and desired applications. While RMSE, MSE, and similar metrics are more widely used in the broader field of modeling, the Pearson correlation coefficient (r) is generally preferred in breeding since knowing the best or worst performing lines is the application. No metric is perfect, Pearson r allows for wildly different absolute values, for example (in fact, 1 submission with absolute values far outside of reasonable corn yields would have scored much better if Pearson r were the metric, see [Supplementary Table 1](#)). The use of Pearson r is recommended for future competitions if the goal is more breeding focused.

To better explore the competition results, numerous metrics were calculated and compared post hoc (see [Supplementary Tables 1 and 2](#)). The scores and rankings based on the average RMSE and other common metrics for the top 20 teams in the competition are shown in [Table 2](#). Rankings that differed from the average RMSE ranking are shaded. Although the absolute ranking of teams changed considerably depending on the metric used, the top and bottom few models were relatively consistent across metrics, with mostly small rank differences observed. The best average RMSE model (the competition winner) remained the winner when evaluated with most other common metrics, although it placed 2nd for global Pearson coefficient (r), a metric that considers accuracy across all environments jointly rather than separately. While the top and bottom teams were fairly robust across metrics, the rank differences seen, which are in some cases large (for example, moving from 15th place to 4th place) illustrate the importance of carefully and deliberately choosing metrics when comparing model accuracies.

Simple average-based ensemble models were also created and tested using combinations of submissions. A model based on all valid submissions combined produced an RMSE of 2.580 and would rank 13th within the leaderboard. A model using the best submission from each of the top 12 teams performed better than the best submission in the competition. This trend continued as the number of top teams decreased with the highest performing ensemble model being a combination of the top 2 teams with an average RMSE score of 2.288. However, the best performing ensemble based on the average of Pearson r environment scores was a model, including the top 5 teams with a score of $r = 0.369$. In summary, some simple ensemble models marginally outperformed the top submissions in the competition.

Some participants focused on predicting environmental means, reasoning that getting those right would be more important than genetics, since the environments were so diverse and RMSE scores favor absolute values over genotypic rankings. The box plots in [Fig. 3a](#) demonstrate that the predictions on average had narrower distributions than the observed values for each environment. Predicting environmental means proved challenging for all teams, particularly for certain environments ([Figs. 3b and 4](#)). For example, WIH1_2022 (Wisconsin) had the highest average yield of all the environments in the test set and was consistently underpredicted by nearly every team ([Supplementary Fig. 2](#)).

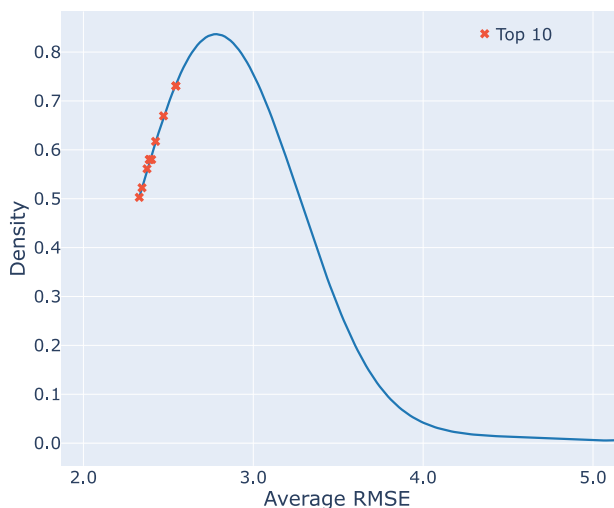


Fig. 2. Density distribution of final competition scores. The top 10 best scores are shown with an "x". The figure is truncated around 5.0 Average RMSE for simplicity.

Table 2. Ranking of the top-20 teams.

Team name	Average of within-environment scores				Across-environment scores			
	RMSE		Pearson r		RMSE		Pearson r	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
CLAC	2.329	1	0.357	1	2.458	1	0.631	2
igorkf	2.345	2	NaN	NaN	2.464	2	0.600	6
phenomaize	2.374	3	0.238	8	2.470	3	0.617	4
UCD_MegaLMM	2.387	4	0.338	3	2.505	6	0.616	5
CGM	2.391	5	0.353	2	2.490	5	0.587	7
Purdue	2.402	6	0.161	15	2.488	4	0.631	3
SmAL	2.425	7	0.146	17	2.525	7	0.586	8
ML_APT	2.472	8	0.191	13	2.600	8	0.564	10
MPB_Group	2.544	9	0.255	6	2.741	11	0.494	13
AlBreeding	2.544	10	0.220	9	2.758	14	0.439	15
AgroStat	2.562	11	0.100	20	2.646	9	0.554	11
arulrich	2.575	12	0.205	12	2.726	10	0.510	12
DataJanitors	2.587	13	0.256	5	2.752	12	0.644	1
CropsAreCool	2.616	14	0.161	14	2.805	15	0.413	19
AllModelsAreWrong	2.646	15	0.272	4	2.754	13	0.575	9
agAdaptAR	2.685	16	0.151	16	2.848	18	0.400	22
CropEnthusiast	2.710	17	0.116	19	2.844	17	0.437	16
DeepCropVision	2.739	18	-0.131	29	2.829	16	0.388	23
supermanwasd	2.746	19	0.243	7	2.915	21	0.409	21
BioSense	2.747	20	0.131	18	2.892	19	0.410	20

Bold values highlight the winning team.

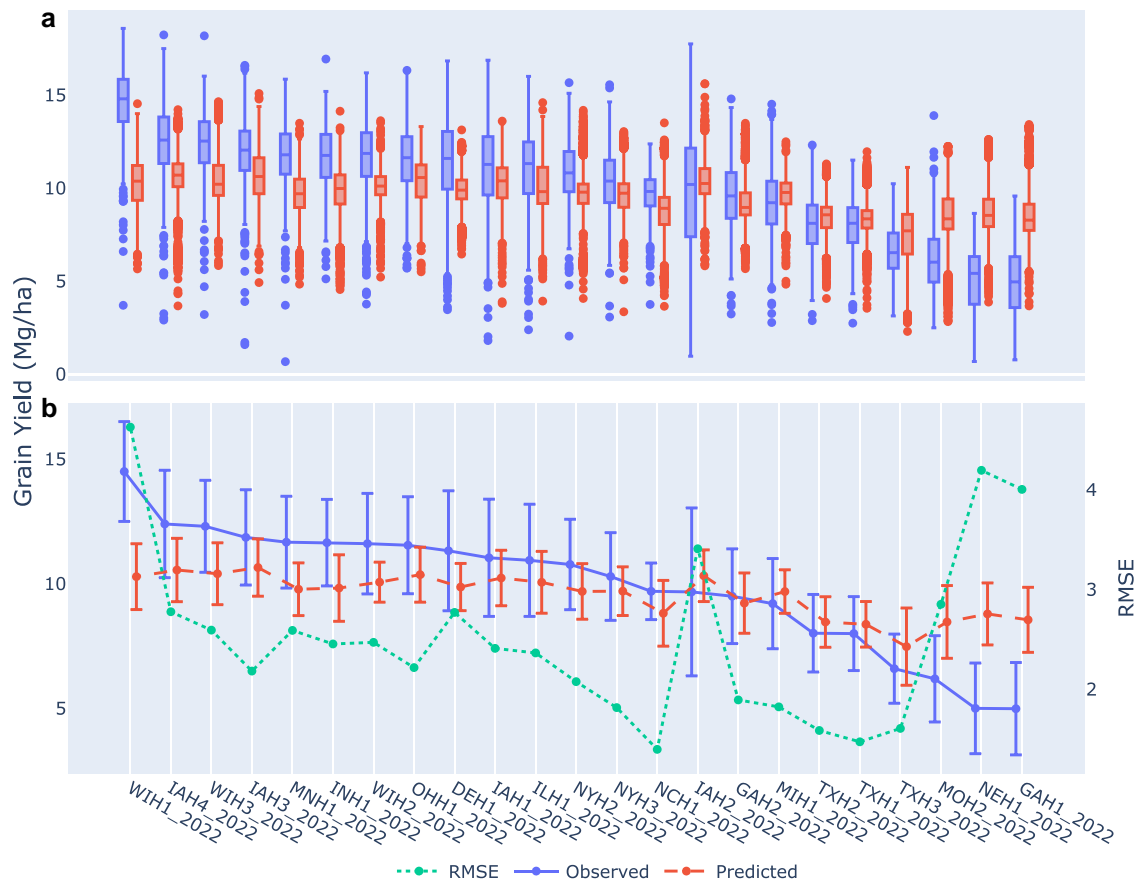


Fig. 3. Test set environmental distributions based on the observed and predicted values from all teams (excluding the final team on the leaderboard due to significant outliers) shown as: a) box plots and b) environment means with standard deviations as error bars and average RMSE scores.

Most midwestern environments were predicted to have similar average yields to one another. Two exceptions were the IAH2 (Iowa), and NEH1 (Nebraska) locations which performed worse

than predicted. IAH2 also had the greatest observed variation of the locations. In contrast, several of the southern and northern locations, including NCH1 (North Carolina) and Texas

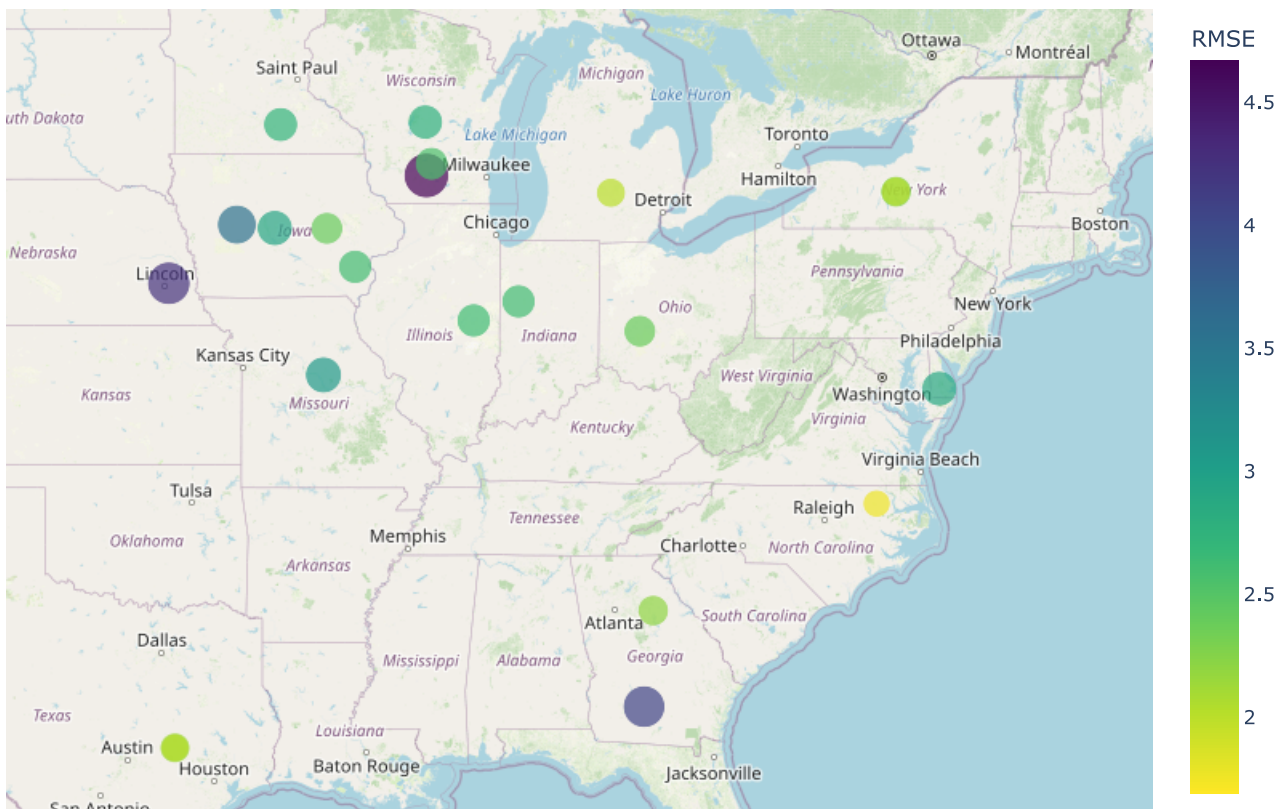


Fig. 4. Average of all team's (excluding the final team on the leaderboard due to significant outliers) per environment RMSE scores plotted on a US map. Both the color and size of the dots represent RMSE scores.

and New York, were predicted very well on average, GAH1 (Georgia) being an exception. In summary, the most difficult environments to predict were those on the extremes with the highest and lowest observed yields and the team's predictions tended toward the mean of all environments (Fig. 3b).

Modeling strategies

Diverse modeling strategies, including different ways of preprocessing the data, were applied by teams in the competition. These approaches included variations on traditional GBLUPs and linear mixed models, deterministic models, RF, ridge regression, gradient-boosting decision trees, deep learning neural networks, and others. Additionally, many teams used combinations of approaches (see [Supplementary File 2](#)). To better understand the approaches used in the competition, a postcompetition survey was sent to all participating teams and 27 of the 30 teams responded. This survey provided broad categorical information about the model types they used. Additionally, for all teams participating in the manuscript, more detailed information on the models used was extracted from each team's methods section. Summaries of the information from both the poll and the methods sections are found in [Supplementary Table 3](#).

Categorizing models based on selected criteria is challenging, as many modeling strategies do not fit discreetly into categories. Some authors have considered modeling methods on a gradient of complexity and/or interpretability ([James et al. 2023](#); [Negus et al. 2024](#)). Although the survey results provided interesting information, the categories required subjective interpretation by the respondents and authors and should be interpreted in that context. Based on the survey, 63% of respondents used models in the "Classical Machine Learning" category which was defined as

excluding simple linear models and deep learning, 52% used simple linear, mixed, or GBLUP-type models, 33% used deep learning methods, 7% used other methods, and 52% used some kind of combination or ensemble approach.

To dive deeper into the different model types used, further model information was collected from the methods of the teams participating in this manuscript. Even this information is somewhat subjective since models can be applied in very different ways, but this provides an interesting overview of the variety of techniques used in the competition. Note that the model types are not exclusive. The majority of teams (76%) used some kind of ensemble method, either as part of their main method(s) (for example, RF uses ensembles), or as an intermediate or final step for combining results from multiple different models/methods. Next, linear models of various kinds (53%), RF models (41%), deep learning (41%), gradient boosting methods (29%), and finally mechanistic/deterministic models (6%) were used. Many variations on these and other methods were used, and specific details can be found in [Supplementary File 2](#) and the code repositories for each team.

Another important decision each team was required to make was which of the data types provided by the competition, and any other public datasets, they would use in their model and how those data would be represented ([Table 3](#)). Genetics was the number one factor used by 93% of respondent teams. There were only 2 teams that did not use genetics in their models, although another team initially excluded genetics from their model but then added it in and found accuracy improvements (see [Supplementary File 2](#)). Surprisingly, one of the teams that explicitly excluded genetic factors ranked in second place in the competition. Importantly though, the ECs provided by the competition and used by this

Table 3. Factors included in the models by team.

Team name	Rank	Factors included in model						
		Genetics	Weather	Soil	Environment covariates	Field management	Experimental design	Other factors
CLAC	1	X	X	X	X	X		
igorkf	2		X	X	X	X		
phenomaize	3	X					X	
UCD_MegaLMM	4	X					X	
CGM	5	X						X
Purdue	6	X	X	X			X	
SmAL	7	X	X	X	X	X		
ML_APT	8	X	X	X	X	X	X	
MPB_Group	9	X	X	X		X		X
AlBreeding	10	X	X	X	X			
AgroStat	11	X	X					
arulrich	12	X	X	X				
DataJanitors	13	X			X	X	X	
CropsAreCool	14		X	X	X	X	X	
AllModelsAreWrong	15	X	X	X	X			X
agAdaptAR	16	X	X	X	X			X
DeepCropVision	18	X	X	X	X	X		X
supermanwasd	19	X	X	X	X			
BioSense	20	X	X	X	X	X		
Kernel of Truth	21	X				X		
AlMaize	24	X	X	X	X			
EnBiSys	25	X	X	X	X	X		
Genetwister	26	X					X	
gartybois	27	X	X	X	X	X	X	
Niche Squad	28	X		X	X			
uwaBioinfo	29	X	X	X				
TinyAfrica	30	X	X	X	X	X		X
Percentage of teams using factor		93%	74%	74%	63%	48%	30%	22%

team were generated by a method that employed some limited use of genetic information (Lopez-Cruz et al. 2023). Additionally, maize hybrids are typically adapted to different climates across the United States, often performing poorly outside of those locations. For this reason, the G2F G×E project applies a stratified approach where not all hybrids are grown in every location and the hybrids grown in any given location are oversampled for those who are adapted to that location’s climate. This likely amounts to some genetics being represented in factors that would otherwise be considered strictly environmental. Additionally, past work on the G2F G×E datasets has indicated substantially more variance among environments than among hybrids, due to the large range of environmental conditions: much larger than a typical hybrid development trial (Rogers et al. 2021; Washburn et al. 2021; Rogers and Holland 2022; Lopez-Cruz et al. 2023). Another important factor is that the competition was based on a very difficult prediction scenario where many of the hybrids in the testing set did not exist in the training set, and even the parents of hybrids in the testing set were largely different from those in the training set. Because of these differences, even the models that explicitly included genetic factors may have struggled to make significant accuracy improvements using them. Regardless of the reasons, this second-place model demonstrates that it is possible to make reasonable predictions based on limited genetic input in the context of the G2F G×E project, even if these predictions would probably not be particularly useful in a breeding context. It is worth noting, as described in the Introduction section, that many important research and application questions can and have been addressed with E/M-centric modeling approaches.

Weather and soil data were the next most highly used factors at 74% each. Along with the ECs, used by 63% of respondents, these

constituted the strictly environmental factors (if human-controlled management is considered separately from nonhuman-controlled environments) provided by the competition. Field management (tillage, irrigation, etc.) and experimental design factors, which include details like field replicates and spatial arrangements in the field, were used by 48 and 30% of respondents, respectively. Additionally, 22% of the teams included other factors, such as external historical datasets, in their models.

Winning team’s strategy and results

The winning team’s strategy involved: (1) identifying and defining the target population of environments (TPEs) and target population of genotypes (TPGs) to better understand the prediction challenge. (2) Considering the implications of the evaluation metric used in the competition (i.e. RMSE favors absolute differences whereas Pearson *r* is a correlation and favors rank differences). (3) Focusing their efforts on modeling location means and genotypic performance separately. The team reasoned that getting the environmental means within a similar scale to the observed values might pay off more than predicting the ranks of genotypes within each environment correctly. Moreover, any modeling artifacts, such as “shrinkage” of coefficients, would be more detrimental to RMSE scores than it might be to Pearson *r* or other rank-based metrics. This type of strategy, focusing on location means, has been suggested and implemented in many studies going back a very long time (Yates and Cochran 1938; Finlay and Wilkinson 1963).

Figure 5 displays the spatial distribution of environments, as year-location combinations, and the prediction targets. All locations in the testing set had already been observed in the previous years of data utilized to calibrate the model (training set),

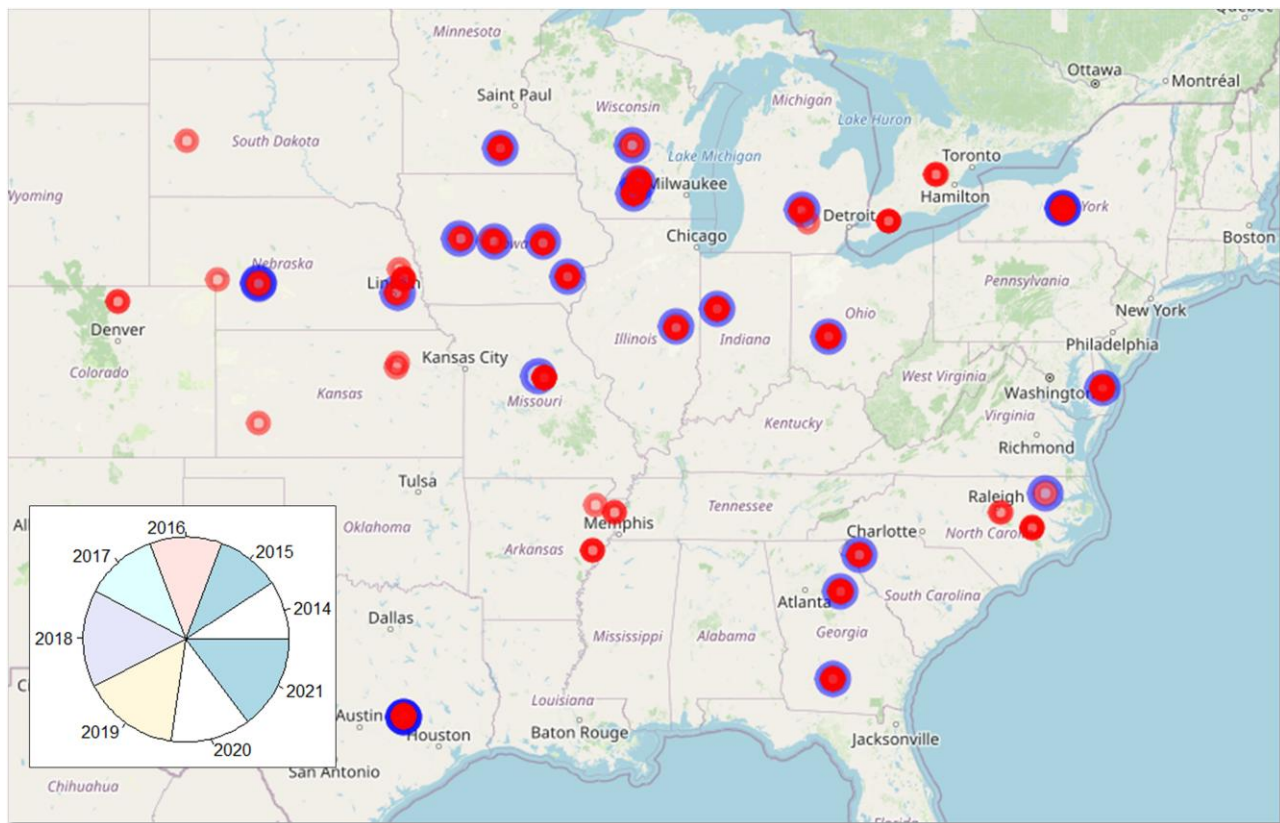


Fig. 5. Target population of environments: distribution of trial data (small circles) and prediction targets (large circles). The pie chart indicates the distribution of data points per year.

indicating that a “location” term could be directly used as a predictor in the model for location means, as well as an informative way to model GxE interactions. Thus, the unobserved environmental means were primarily informed by the average location performance from previous years, complemented by the available information in the provided metadata (e.g. treatment, previous crop, etc.) treated as covariates, and ECs as captured by an RF model.

Regarding the TPGs, Fig. 6 indicates that the prediction target is contained within the genotypic space of the training data based on the 2 first principal components, although it is not necessarily well represented. Genomic information was provided at the hybrid level, and the clouds seem to indicate substantial contribution of testers to the population structure. Relationship information was utilized by the winning team to infer the determinist accuracy between every pair of locations, specifically between observed and target environments. Together, deterministic accuracy estimates from genomic information harnessing the TPG, and the spatial location, capturing the TPEs were used to assign the contribution of each training environment to the prediction of each testing environment. Since the shrinkage of genomic values was likely to have a detrimental impact on the evaluation metric, predicted genomic values (from marker information) were normalized and rescaled. This was done using a standard deviation value the team predicted for each environment based on a model analogous to the location mean model.

Each of the winning team’s submissions is listed in Table 4. The modifications made between each submission were based on the feedback (RMSE scores) from the previous submission and conjectures made by the team. To gain degrees of freedom and

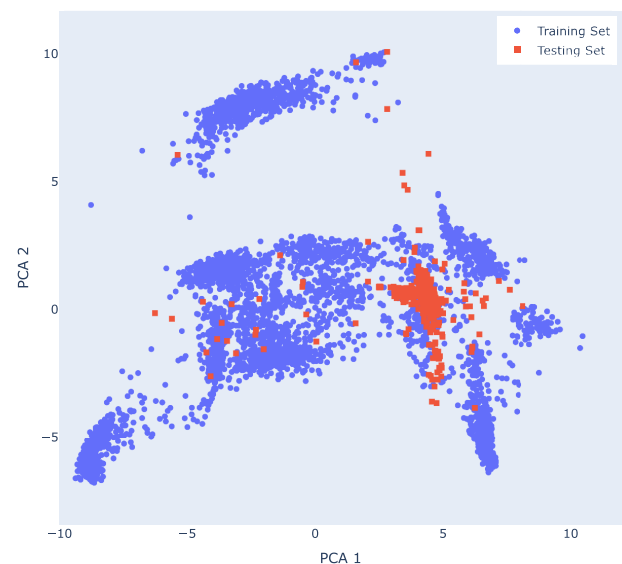


Fig. 6. Target population of genotypes: first and second principal components of the genomic relationship matrix with testing set genotypes (red dots) and training set genotypes (blue dots).

predictive power, the data concerning previous crops were aggregated into 4 levels: wheat, legume, corn, or other. Treatments were also aggregated into standard, dry, and late. Irrigation information (yes/no) was removed from treatments, and it was attributed to an additional covariate.

The team’s first submission was based on their best guess and reasoned strategies from prior experience: a multivariate model with minimal data QC. With the hope of improving this model for their second submission, members of the team with breeding knowledge reviewed the data and applied more stringent QC. This involved dropping outliers using a 3-standard deviation rule, removing locations with treatments tagged as “disease trial,” and removing experimental units with stand counts below 20. However, this second submission with greater QC performed worse than the first one, possibly because the quality of the testing set was comparable with the training data, causing the data QC to make the dataset less representative of the prediction target. At this point, the team was uncertain how to proceed but determined to submit their first model averaged with a slightly simpler model without any ECs. This resulted in a better RMSE than either of the first 2 submissions. To determine whether the simpler model was better by itself, they submitted it alone as their fourth attempt. The poorer result indicated it was not. For their final submission, the team used their original model (from submission one) but averaged it with an even simpler univariate GBLUP than submission 4, resulting in the winning score. Although the winning team’s strategy was in part focused on predicting environmental means, their final winning submission actually resulted in a large increase in Pearson r over previous submissions (Fig. 7), indicating that better within-environment (genetic) prediction was a component of what boosted them to the winning position. After further postcompetition study, they concluded that because the data were very structured, most of the predictive ability within-environment was probably coming from the univariate GBLUP rather than the unstructured/multivariate GxE model.

Conclusions

A large diversity of model types, strategies, and data inputs were used in the competition and, perhaps surprisingly, many of these made it into the top 10 list. This demonstrates that no single modeling strategy is greatly superior to the others, at least in the context of this prediction problem. Additionally, ensemble approaches, even those as simple as averaging, combining multiple models with different strengths and weakness often outperform single models. The winning team’s strategy of attempting to model environmental means separately from genetics appears to have been effective with the simple univariate GBLUP from the final model improving both absolute (RMSE) and relative/rank-based (Pearson r) scores within most environments.

Future competitions might consider predicting additional traits beside yield, since different methods might be impacted differently by the genetic architecture or other factors associated with a given trait. Different traits can also have different GxE interactions. Factors other than accuracy scores could also be considered, for example, the computational requirements and speed of training. However, these factors can be difficult to track in a fair and objective manner, and determining which factors are most important is not necessarily strait forward. Additionally, the computational resources commonly available are changing (i.e. very powerful graphics cards can be used for deep learning methods and are readily available today, even to the general public).

The level of international interest and participation in the competition demonstrates that there is broad excitement about crop yield prediction from many disciplines and carrier stages. Crop yield prediction comes with both significant challenges and enormous potential benefits to society. Although the bulk of registrants came from plant science, genetics, or breeding

Table 4. Winning team’s modeling strategy by submission.

Submission	RMSE	r	Model description
S1	2.454	0.263	GBLUP with minor QC and some ECs
S2	2.521	0.260	GBLUP with QC on GEBVs and location means; no ECs
S3	2.353	0.266	Ensemble: average of methods used in S1 and S4
S4	2.410	0.267	GBLUP with QC on GEBVs; no location means QC; no ECs
S5	2.329	0.357	Ensemble: average of S1 and univariate GBLUP

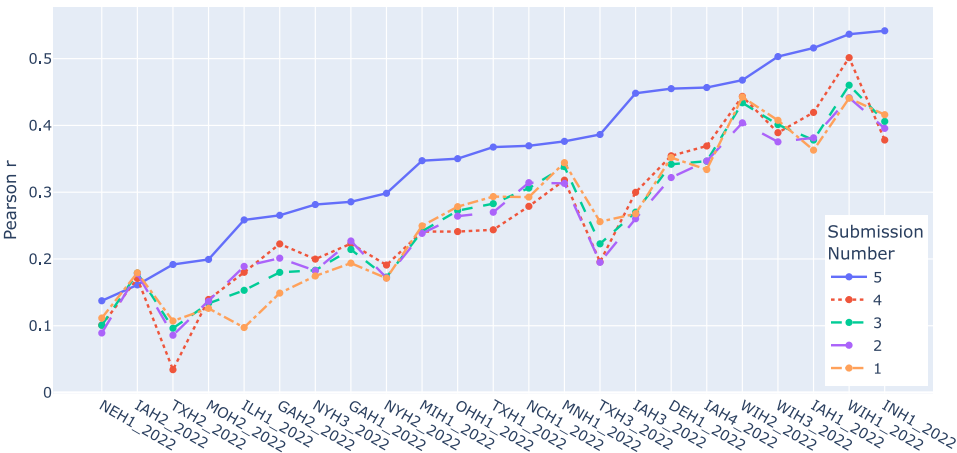


Fig. 7. Pearson r values from the winning team by location and submission.

backgrounds, participants with little or no background in these areas also participated. Activities like this competition have the potential to bring new ideas and ways of thinking from disciplines like computer science and engineering into the genetics and plant science communities, enhancing our abilities to solve critical, and technically challenging, problems.

Data availability

All data used in this manuscript are publicly available at <https://doi.org/10.25739/tq5e-ak26>. Code from all teams is publicly available as follows:

AgAdaptAR: <https://github.com/EcoEvoInfo/maize-gxe-prediction-challenge-2023>

AIMaize: https://github.com/ksegaba/Genomes2Field_Competition

All Models are Wrong: <https://zenodo.org/record/7830071>

arulrich: <https://github.com/mwylerCH/GxEcompetition>

CLAC: https://github.com/alenvxav/Lectures/tree/master/MGC_2023

DataJanitors: <https://github.com/qchen33/g2fcompetition2022>

DeepCropVision: https://github.com/Ved-Piyush/DeepCropVision_maizexprediction2022

EnBiSys: <https://github.com/dperondi/maizexprediction2022>

gartybois: <https://github.com/Thyra/g2f-maize-challenge-2022>

Kernel of Truth: https://github.com/robertkhu/maize_gxe

ML_APT: https://forgemia.inra.fr/ml_apt/g2f_challenge

MPB_Group: <https://zenodo.org/records/12721443>

Niche Squad: <https://github.com/Niche-Squad/gsfomer>

Phenomaize: <https://forgemia.inra.fr/renaud.rincent/genome2fields>

SmAL: <https://github.com/SmartAgriLabs/G2F-competition>

UCD_MegaLMM: https://github.com/ucdavis/UCD_MegaLMM

uwaBioinfo: <https://github.com/eyesoftruth/G2F-competition2022>

Supplemental material available at GENETICS online.

Acknowledgments

The authors thank the field managers, crop research coordinators, staff, graduate students, student interns, and data collectors for their efforts in the GxE project. Team phenomaize is grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi:10.15454/1.5572390655343 293E12) for providing computing resources.

Funding

J.L.G.: This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Institute of Food and Agriculture (7002327). Research reported in this publication was supported by the National Institutes of Health, National Institute of General Medical Sciences under award number R35GM151048. Support also came from the U.S. Department of Agriculture, Agricultural Research Service, Iowa Corn Growers Association, and National Corn Growers Association. D.R.K.: This research was supported by the U. S. Department of Agriculture, Agricultural Research Service (project number 5070-21000-041-000-D). D.E.R. and H.H.: This work was supported by the Agriculture and Food Research Initiative grant no. 2020-67013-30904 from the U.S. Department of Agriculture, National Institute of Food and Agriculture. The

generation of ECs was funded by National Science Foundation Plant Genome Research Program grant #2035472.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

- Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M et al. 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv 1603.04467. <https://doi.org/10.48550/arXiv.1603.04467>.
- Anche MT, Kaczmar NS, Morales N, Clohessy JW, Ilut DC, Gore MA, Robbins KR. 2020. Temporal covariance structure of multi-spectral phenotypes and their predictive ability for end-of-season traits in maize. *Theor Appl Genet.* 133(10):2853–2868. doi:10.1007/s00122-020-03637-6.
- Anderson II SL, Murray SC, Malambo L, Ratcliff C, Popescu S, Cope D, Chang A, Jung J, Thomasson JA. 2019. Prediction of maize grain yield before maturity using improved temporal height estimates of unmanned aerial systems. *Plant Phenome J.* 2(1):190004. doi:10.2135/tppj2019.02.0004.
- Archontoulis SV, Miguez FE, Moore KJ. 2014. A methodology and an optimization tool to calibrate phenology of short-day species included in the APSIM plant model: application to soybean. *Environ Model Softw.* 62:465–477. doi:10.1016/j.envsoft.2014.04.009.
- Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H. 2019. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 (Bethesda).* 9(11):3691–3702. doi:10.1534/g3.119.400498.
- Bai G, Ge Y, Scoby D, Leavitt B, Stoerger V, Kirchgessner N, Irmak S, Graef G, Schnable J, Awada T. 2019. NU-Spidercam: a large-scale, cable-driven, integrated sensing and robotic system for advanced phenotyping, remote sensing, and agronomic research. *Comput Electron Agric.* 160:71–81. doi:10.1016/j.compag.2019.03.009.
- Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, et al. 2016. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. 7:221. doi:10.3389/fgene.2016.00221.
- Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. 2022. Deep neural networks and tabular data: a survey. *IEEE Trans Neural Netw Learn Syst.* 35(6):7499–7519. doi:10.1109/TNNLS.2022.3229161.
- Bornowski N, Michel KJ, Hamilton JP, Ou S, Seetharam AS, Jenkins J, Grimwood J, Plott C, Shu S, Talag J, et al. 2021. Genomic variation within the maize stiff-stalk heterotic germplasm pool. *Plant Genome.* 14(3):e20114. doi:10.1002/tpg2.20114.
- Bradbury PJ, Casstevens T, Jensen SE, Johnson LC, Miller ZR, Monier B, Romay MC, Song B, Buckler ES. 2022. The practical haplotype graph, a platform for storing and using pangenomes for imputation. *Bioinformatics.* 38(15):3698–3702. doi:10.1093/bioinformatics/btac410.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics.* 23(19):2633–2635. doi:10.1093/bioinformatics/btm308.
- Breiman L. 2001. Random forests. *Mach Learn.* 45(1):5–32. doi:10.1023/A:1010933404324.
- Budhlakoti N, Kushwaha AK, Rai A, Chaturvedi KK, Kumar A, Pradhan AK, Kumar U, Kumar RR, Juliana P, Mishra DC, et al.

2022. Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Front Genet.* 13:832153. doi:[10.3389/fgene.2022.832153](https://doi.org/10.3389/fgene.2022.832153).
- Butler DG, Cullis BR, Gilmour AR, Gogel BG, Thompson R. 2017. ASReml-R Reference Manual Version 4. Hemel Hempstead (UK): VSN International Ltd.
- Challinor AJ, Müller C, Asseng S, Deva C, Nicklin KJ, Wallach D, Vanuytrecht E, Whitfield S, Ramirez-Villegas J, Koehler A-K. 2018. Improving the use of crop models for risk assessment and climate change adaptation. *Agric Syst.* 159:296–306. doi:[10.1016/j.agry.2017.07.010](https://doi.org/10.1016/j.agry.2017.07.010).
- Charmet G, Tran L-G, Auzanneau J, Rincet R, Bouchet S. 2020. BWGS: a R package for genomic selection and its application to a wheat breeding programme. *PLoS One.* 15(4):e0222733. doi:[10.1371/journal.pone.0222733](https://doi.org/10.1371/journal.pone.0222733).
- Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* p. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Chollet F. Keras. <https://github.com/fchollet/keras>. Accessed January 15 20232015.
- Cooper M, Messina CD, Tang T, Gho C, Powell OM, Podlich DW, Technow F, Hammer GL. 2022. Predicting genotype × environment × management (g × e × m) interactions for the design of crop improvement strategies. In: Goldman I, editor. *Plant Breeding Reviews*. Vol. 46. Hoboken (NJ): Wiley. p. 467–585.
- Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, et al. 2017. Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22(11):961–975. doi:[10.1016/j.tplants.2017.08.011](https://doi.org/10.1016/j.tplants.2017.08.011).
- Cuevas J, Montesinos-López O, Juliana P, Guzmán C, Pérez-Rodríguez P, González-Bucio J, Burgueño J, Montesinos-López A, Crossa J. 2019. Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 (Bethesda).* 9(9):2913–2924. doi:[10.1534/g3.119.400493](https://doi.org/10.1534/g3.119.400493).
- Cvejosi K, Schuecker J, Mahlein A-K, Georgiev B. 2021. Combining expert knowledge and neural networks to model environmental stresses in agriculture. *arXiv* 2111.00918. <https://doi.org/10.48550/arXiv.2111.00918>.
- Danilevicz MF, Bayer PE, Boussaid F, Bennamoun M, Edwards D. 2021. Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sens.* 13(19):3976. doi:[10.3390/rs13193976](https://doi.org/10.3390/rs13193976).
- DeChant C, Wiesner-Hanks T, Chen S, Stewart EL, Yosinski J, Gore MA, Nelson RJ, Lipson H. 2017. Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. *Phytopathology.* 107(11):1426–1432. doi:[10.1094/PHYTO-11-16-0417-R](https://doi.org/10.1094/PHYTO-11-16-0417-R).
- Desta ZA, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19(9):592–601. doi:[10.1016/j.tplants.2014.05.006](https://doi.org/10.1016/j.tplants.2014.05.006).
- Diepenbrock CH, Tang T, Jines M, Technow F, Lira S, Podlich D, Cooper M, Messina C. 2022. Can we harness digital technologies and physiology to hasten genetic gain in US maize breeding? *Plant Physiol.* 188(2):1141–1157. doi:[10.1093/plphys/kiab527](https://doi.org/10.1093/plphys/kiab527).
- Di Paola A, Valentini R, Santini M. 2016. An overview of available crop growth and yield models for studies and assessments in agriculture. *J Sci Food Agric.* 96(3):709–714. doi:[10.1002/jsfa.7359](https://doi.org/10.1002/jsfa.7359).
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One.* 6(5):e19379. doi:[10.1371/journal.pone.0019379](https://doi.org/10.1371/journal.pone.0019379).
- Falcon CM, Kaeppler SM, Spalding EP, Miller ND, Haase N, Alkhalifah N, Bohn M, Buckler ES, Campbell DA, Ciampitti I, et al. 2020. Relative utility of agronomic, phenological, and morphological traits for assessing genotype-by-environment interaction in maize inbreds. *Crop Sci.* 60(1):62–81. doi:[10.1002/csc2.20035](https://doi.org/10.1002/csc2.20035).
- Fernandes IK, Vieira CC, Dias KOG, Fernandes SB. 2024. Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials. *Theor Appl Genet.* 137(8):189. doi:[10.1007/s00122-024-04687-w](https://doi.org/10.1007/s00122-024-04687-w).
- Finlay KW, Wilkinson GN. 1963. The analysis of adaptation in a plant-breeding programme. *Aust J Agric Res.* 14:742–754. doi:[10.1071/AR9630742](https://doi.org/10.1071/AR9630742).
- Gage JL, Jarquin D, Romay C, Lorenz A, Buckler ES, Kaeppler S, Alkhalifah N, Bohn M, Campbell DA, Edwards J, et al. 2017. The effect of artificial selection on phenotypic plasticity in maize. *Nat Commun.* 8(1):1348. doi:[10.1038/s41467-017-01450-2](https://doi.org/10.1038/s41467-017-01450-2).
- Gage JL, Richards E, Lepak N, Kaczmar N, Soman C, Chowdhary G, Gore MA, Buckler ES. 2019. In-field whole-plant maize architecture characterized by subcanopy rovers and latent space phenotyping. *Plant Phenome J.* 2(1):190011. doi:[10.2135/tppj2019.07.0011](https://doi.org/10.2135/tppj2019.07.0011).
- Ge Z, Zhou Z, Xuye K, Yitong C. 2024. FF-LSTM: phenotype prediction based on feature fusion. In *ProcSPE. 3rd International Conference on Electronic Information Engineering and Data Processing*, Kuala Lumpur, Malaysia. SPIE Digital Library. p. 13184:131846H.
- Gill M, Anderson R, Hu H, Bennamoun M, Petereit J, Valliyodan B, Nguyen HT, Batley J, Bayer PE, Edwards D. 2022. Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biol.* 22(1):180. doi:[10.1186/s12870-022-03559-z](https://doi.org/10.1186/s12870-022-03559-z).
- González-Recio O, Forni S. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genet Sel Evol.* 43(1):7. doi:[10.1186/1297-9686-43-7](https://doi.org/10.1186/1297-9686-43-7).
- Grinstajn L, Oyallon E, Varoquaux G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv* 2207.08815. <https://doi.org/10.48550/arXiv.2207.08815>.
- Guo T, Li X. 2023. Machine learning for predicting phenotype from genotype and environment. *Curr Opin Biotechnol.* 79:102853. doi:[10.1016/j.copbio.2022.102853](https://doi.org/10.1016/j.copbio.2022.102853).
- Haley CS, Visscher PM. 1998. Strategies to utilize marker-quantitative trait loci associations. *J Dairy Sci.* 81:85–97. doi:[10.3168/jds.S0022-0302\(98\)70157-2](https://doi.org/10.3168/jds.S0022-0302(98)70157-2).
- Hammer GL, Kropff MJ, Sinclair TR, Porter JR. 2002. Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. *Eur J Agron.* 18(1):15–31. doi:[10.1016/S1161-0301\(02\)00093-X](https://doi.org/10.1016/S1161-0301(02)00093-X).
- Hammer G, McLean G, Doherty A, van Oosterom E, Chapman S. 2019. Sorghum Crop Modeling and its Utility in Agronomy and Breeding. In: Ciampitti IA, Vara Prasad PV, editors. *Sorghum: A State of the Art and Future Perspectives*. American Society of Agronomy Crop Science Society of America Soil Science Society of America. p. 215–239.
- Heffner EL, Sorrells ME, Jannink J-L. 2009. Genomic selection for crop improvement. *Crop Sci.* 49(1):1–12. doi:[10.2135/cropsci2008.08.0512](https://doi.org/10.2135/cropsci2008.08.0512).
- Hesami M, Alizadeh M, Naderi R, Tohidfar M. 2020. Forecasting and optimizing agrobacterium-mediated genetic transformation via ensemble model- fruit fly optimization algorithm: a data mining approach using chrysanthemum databases. *PLoS One.* 15(9):e0239901. doi:[10.1371/journal.pone.0239901](https://doi.org/10.1371/journal.pone.0239901).
- Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. 2021. De novo assembly,

- annotation, and comparative analysis of 26 diverse maize genomes. *Science*. 373(6555):655–662. doi:[10.1126/science.abg5289](https://doi.org/10.1126/science.abg5289).
- James G, Witten D, Hastie T, Tibshirani R, Taylor J. 2023. *An Introduction to Statistical Learning with Applications in Python*. Switzerland: Springer Cham.
- Jarquín D, de Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, et al. 2021. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front Genet*. 11:592769. doi:[10.3389/fgene.2020.592769](https://doi.org/10.3389/fgene.2020.592769).
- 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(3):595–607. doi:[10.1007/s00122-013-2243-1](https://doi.org/10.1007/s00122-013-2243-1).
- Johnsson M. 2023. Genomics in animal breeding from the perspectives of matrices and molecules. *Hereditas*. 160(1):20. doi:[10.1186/s41065-023-00285-w](https://doi.org/10.1186/s41065-023-00285-w).
- Jones JW, Hoogenboom G, Porter CH, Boote KJ, Batchelor WD, Hunt LA, Wilkens PW, Singh U, Gijsman AJ, Ritchie JT. 2003. The dssat cropping system model. *Eur J Agron*. 18(3):235–265. doi:[10.1016/S1161-0301\(02\)00107-7](https://doi.org/10.1016/S1161-0301(02)00107-7).
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. 2017. LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach (CA, USA): Curran Associates Inc.
- Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, Huth NI, Hargreaves JNG, Meinke H, Hochman Z, et al. 2003. An overview of APSIM, a model designed for farming systems simulation. *Eur J Agron*. 18(3):267–288. doi:[10.1016/S1161-0301\(02\)00108-9](https://doi.org/10.1016/S1161-0301(02)00108-9).
- Khaki S, Wang L. 2019. Crop yield prediction using deep neural networks. *Front Sci*. 10:621. doi:[10.3389/fpls.2019.00621](https://doi.org/10.3389/fpls.2019.00621).
- Khalilzadeh Z, Sajid SS, Khaki S, Wang L, Hu G. 2024. Comprehensive Crop Yield Prediction Using Transformer-Enhanced Neural Networks Considering Different Combinations of Sequential Data Including Weather, Genotype, and APSIM Datasets and Non-Sequential Data. Ames (IA): Iowa State University.
- Kick DR, Wallace JG, Schnable JC, Kolkman JM, Alaca B, Beissinger TM, Edwards J, Ertl D, Flint-Garcia S, Gage JL, et al. 2023. Yield prediction through integration of genetic, environment, and management data through deep learning. *G3 (Bethesda)*. 13(4):jkad006. doi:[10.1093/g3journal/jkad006](https://doi.org/10.1093/g3journal/jkad006).
- Kick DR, Washburn JD. 2023. Ensemble of best linear unbiased predictor, machine learning and deep learning models predict maize yield better than each model alone. *in silico Plants*. 5(2):diad015. doi:[10.1093/insilicoplants/diad015](https://doi.org/10.1093/insilicoplants/diad015).
- Li X, Guo T, Wang J, Bekele WA, Sukumaran S, Vanous AE, McNellie JP, Tibbs-Cortes LE, Lopes MS, Lamkey KR, et al. 2021. An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. *Mol Plant*. 14(6):874–887. doi:[10.1016/j.molp.2021.03.010](https://doi.org/10.1016/j.molp.2021.03.010).
- Liang M, Miao J, Wang X, Chang T, An B, Duan X, Xu L, Gao X, Zhang L, Li J, et al. 2021. Application of ensemble learning to genomic selection in Chinese Simmental beef cattle. *J Anim Breed Genet*. 138(3):291–299. doi:[10.1111/jbg.12514](https://doi.org/10.1111/jbg.12514).
- Lima DC, Aviles AC, Alpers RT, McFarland BA, Kaeppler S, Ertl D, Romay MC, Gage JL, Holland J, Beissinger T, et al. 2023a. 2018–2019 Field seasons of the maize genomes to fields (g2f) g x e project. *BMC Genom Data*. 24(1):29. doi:[10.1186/s12863-023-01129-2](https://doi.org/10.1186/s12863-023-01129-2).
- Lima DC, Aviles AC, Alpers RT, Perkins A, Schoemaker DL, Costa M, Michel KJ, Kaeppler S, Ertl D, Romay MC, et al. 2023b. 2020–2021 Field seasons of maize GxE project within the genomes to fields initiative. *BMC Res Notes*. 16(1):219. doi:[10.1186/s13104-023-06430-y](https://doi.org/10.1186/s13104-023-06430-y).
- Lima DC, Washburn JD, Varela JI, Chen Q, Gage JL, Romay MC, Holland J, Ertl D, Lopez-Cruz M, Aguade FM, et al. 2023c. Genomes to fields 2022 maize genotype by environment prediction competition. *BMC Res Notes*. 16(1):148. doi:[10.1186/s13104-023-06421-z](https://doi.org/10.1186/s13104-023-06421-z).
- Lin Z, Robinson H, Godoy J, Rattey A, Moody D, Mullan D, Keeble-Gagnere G, Forrest K, Tibbits J, Hayden MJ, et al. 2021. Genomic prediction for grain yield in a barley breeding program using genotype x environment interaction clusters. *Crop Sci*. 61(4):2323–2335. doi:[10.1002/csc2.20460](https://doi.org/10.1002/csc2.20460).
- Lopez-Cruz M, Aguade FM, Washburn JD, de Leon N, Kaeppler SM, Lima DC, Tan R, Thompson A, De La Bretonne LW, de los Campos G. 2023. Leveraging data from the genomes-to-fields initiative to investigate genotype-by-environment interactions in maize in North America. *Nat Commun*. 14(1):6904. doi:[10.1038/s41467-023-42687-4](https://doi.org/10.1038/s41467-023-42687-4).
- Lopez-Cruz M, Pérez-Rodríguez P, de los Campos G. 2024. A fast algorithm to factorize high-dimensional tensor product matrices used in genetic models. *G3 (Bethesda)*. 14(3):jkae001. doi:[10.1093/g3journal/jkae001](https://doi.org/10.1093/g3journal/jkae001).
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L. 2011. Chapter 2: Genomic selection in plant breeding: knowledge and prospects. In: Sparks DL, editor. *Advances in Agronomy*. Academic Press. p. 77–123.
- Ly D, Chenu K, Gauffreteau A, Rincet R, Huet S, Gouache D, Martre P, Bordes J, Charmet G. 2017. Nitrogen nutrition index predicted by a crop model improves the genomic prediction of grain number for a bread wheat core collection. *Field Crops Res*. 214:331–340. doi:[10.1016/j.fcr.2017.09.024](https://doi.org/10.1016/j.fcr.2017.09.024).
- Malhotra P, Vig L, Shroff G, Agarwal P. 2015. Long short term memory networks for anomaly detection in time series. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015; 2015 Apr 22–24; Bruges (Belgium)*. CIACO sc.
- Martinez ND. 2023. Predicting ecosystem metaphenome from community metagenome: a grand challenge for environmental biology. *Ecol Evol*. 13(3):e9872. doi:[10.1002/ece3.9872](https://doi.org/10.1002/ece3.9872).
- McFarland BA, AlKhalifah N, Bohn M, Buber J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, et al. 2020. Maize genomes to fields (g2f): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Res Notes*. 13(1):71. doi:[10.1186/s13104-020-4922-8](https://doi.org/10.1186/s13104-020-4922-8).
- Messina CD, Gho C, Hammer GL, Tang T, Cooper M. 2023. Two decades of harnessing standing genetic variation for physiological traits to improve drought tolerance in maize. *J Exp Bot*. 74(16):4847–4861. doi:[10.1093/jxb/erad231](https://doi.org/10.1093/jxb/erad231).
- Messina CD, Technow F, Tang T, Totir R, Gho C, 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. doi:[10.1016/j.eja.2018.01.007](https://doi.org/10.1016/j.eja.2018.01.007).
- Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157(4):1819–1829. doi:[10.1093/genetics/157.4.1819](https://doi.org/10.1093/genetics/157.4.1819).
- Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F. 2019. Genomic prediction of maize yield across European environmental conditions. *Nat Genet*. 51(6):952–956. doi:[10.1038/s41588-019-0414-y](https://doi.org/10.1038/s41588-019-0414-y).

- Möhring J, Piepho HP. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Sci.* 49(6):1977–1988. doi:10.2135/cropsci2009.02.0083.
- Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J. 2021. A review of deep learning applications for genomic selection. *BMC Genomics.* 22(1):19. doi:10.1186/s12864-020-07319-x.
- Montesinos López OA, Montesinos López A, Crossa J. 2022. Random forest for genomic prediction. In: Montesinos López OA, Montesinos López A, Crossa J, editors. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer International Publishing. p. 633–681.
- Morales N, Kaczmar NS, Santantonio N, Gore MA, Mueller LA, Robbins KR. 2020. Imagebreed: open-access plant breeding web-database for image-based phenotyping. *Plant Phenome J.* 3(1):e20004. doi:10.1002/ppj2.20004.
- Morota G, Gianola D. 2014. Kernel-based whole-genome prediction of complex traits: a review. *Front Genet.* 5:363. doi:10.3389/fgene.2014.00363.
- National Research Council (US). 2010. *Research at the Intersection of the Physical and Life Sciences*. National Academies Press (US).
- Negus KL, Li X, Welch SM, Yu J. 2024. Chapter 1: The role of artificial intelligence in crop improvement. In: Sparks DL, editor. *Advances in Agronomy*. Academic Press. p. 1–66.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al. 2019. Pytorch: an imperative style, high-performance deep learning library. *arXiv 1912.01703*. <https://doi.org/10.48550/arXiv.1912.01703>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in python. *J Mach Learn Res.* 12: 2825–2830.
- Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics.* 198(2):483–495. doi:10.1534/genetics.114.164442.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna (Austria): R Foundation for Statistical Computing.
- Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, Ertl D, Flint-Garcia S, Gore MA, et al. 2021. The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 (Bethesda).* 11(2):jkac050. doi:10.1093/g3journal/jkaa050.
- Rogers AR, Holland JB. 2022. Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3 (Bethesda).* 12(2):jkab440. doi:10.1093/g3journal/jkab440.
- Schauberger B, Archontoulis S, Armeth A, Balkovic J, Ciaia P, Deryng D, Elliott J, Folberth C, Khabarov N, Müller C, et al. 2017. Consistent negative response of us crops to high temperatures in observations and crop models. *Nat Commun.* 8(1):13931. doi:10.1038/ncomms13931.
- Sekhon RS, Joyner CN, Ackerman AJ, McMahan CS, Cook DD, Robertson DJ. 2020. Stalk bending strength is strongly associated with maize stalk lodging incidence across multiple environments. *Field Crops Res.* 249:107737. doi:10.1016/j.fcr.2020.107737.
- Shahhosseini M, Hu G, Archontoulis SV. 2020. Forecasting corn yield with machine learning ensembles. *Front Plant Sci.* 11:1120. doi:10.3389/fpls.2020.01120.
- Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK. 2021. Crop yield prediction integrating genotype and weather variables using deep learning. *PLoS One.* 16(6): e0252402. doi:10.1371/journal.pone.0252402.
- Shwartz-Ziv R, Armon A. 2022. Tabular data: deep learning is not all you need. *Inf Fusion.* 81:84–90. doi:10.1016/j.inffus.2021.11.011.
- Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. Anchorwave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc Natl Acad Sci U S A.* 119(1):e2113075119. doi:10.1073/pnas.2113075119.
- Stewart EL, Wiesner-Hanks T, Kaczmar N, DeChant C, Wu H, Lipson H, Nelson RJ, Gore MA. 2019. Quantitative phenotyping of northern leaf blight in UAV images using deep learning. *Remote Sens.* 11(19):2209. doi:10.3390/rs11192209.
- Technow F, Messina CD, Totir LR, Cooper M. 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PLoS One.* 10(6):e0130855. doi:10.1371/journal.pone.0130855.
- Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, et al. 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics.* 15(1):823. doi:10.1186/1471-2164-15-823.
- US National Science Foundation. 2023. NSF's 10 big ideas: Understanding the rules of life. U.S. National Science Foundation; [cited 2023 Dec 13]. Available from https://www.nsf.gov/news/special_reports/big_ideas/life.jsp.
- Van Rossum G, Drake FL. 2009. *Python 3 Reference Manual*. CreateSpace.
- Vitezica ZG, Varona L, Legarra A. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics.* 195(4):1223–1230. doi:10.1534/genetics.113.155176.
- Washburn JD, Burch MB, Franco JAV. 2020. Predictive breeding for maize: making use of molecular phenotypes, machine learning, and physiological crop models. *Crop Sci.* 60(2):622–638. doi:10.1002/csc2.20052.
- Washburn JD, Cimen E, Ramstein G, Reeves T, O'Brian P, McLean G, Cooper M, Hammer G, Buckler ES. 2021. Predicting phenotypes from genetic, environment, management, and historical data using CNNs. *Theor Appl Genet.* 134(12):3997–4011. doi:10.1007/s00122-021-03943-7.
- Westhues CC, Mahone GS, da Silva S, Thorwarth P, Schmidt M, Richter J-C, Simianer H, Beissinger TM. 2021. Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Front Plant Sci.* 12:699589. doi:10.3389/fpls.2021.699589.
- Westhues CC, Simianer H, Beissinger TM. 2022. LearnMET: an r package to apply machine learning methods for genomic prediction using multi-environment trial data. *G3 (Bethesda).* 12(11):jkac226. doi:10.1093/g3journal/jkac226.
- Wiesner-Hanks T, Stewart EL, Kaczmar N, DeChant C, Wu H, Nelson RJ, Lipson H, Gore MA. 2018. Image set for deep learning: field images of maize annotated with disease symptoms. *BMC Res Notes.* 11(1):440. doi:10.1186/s13104-018-3548-6.
- Wiesner-Hanks T, Wu H, Stewart E, DeChant C, Kaczmar N, Lipson H, Gore MA, Nelson RJ. 2019. Millimeter-level plant disease detection from aerial photographs via deep learning and crowdsourced data. *Front Plant Sci.* 10:1550. doi:10.3389/fpls.2019.01550.
- Wiggins GR, Cole JB, Hubbard SM, Sonstegard TS. 2017. Genomic selection in dairy cattle: the USDA experience. *Annu Rev Anim Biosci.* 5(1):309–327. doi:10.1146/annurev-animal-021815-111422.
- Winn CA, Archontoulis S, Edwards J. 2023. Calibration of a crop growth model in APSIM for 15 publicly available corn hybrids in North America. *Crop Sci.* 63(2):511–534. doi:10.1002/csc2.20857.

- Woodhouse MR, Cannon EK, Portwood JL, Harper LC, Gardiner JM, Schaeffer ML, Andorf CM. 2021. A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol.* 21(1):385. doi:[10.1186/s12870-021-03173-5](https://doi.org/10.1186/s12870-021-03173-5).
- Wright MN, Ziegler A. 2017. Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *J Stat Softw.* 77(1):1–17. doi:[10.18637/jss.v077.i01](https://doi.org/10.18637/jss.v077.i01).
- Wu H, Wiesner-Hanks T, Stewart EL, DeChant C, Kaczmar N, Gore MA, Nelson RJ, Lipson H. 2019. Autonomous detection of plant disease symptoms directly from aerial imagery. *Plant Phenome J.* 2(1):190006. doi:[10.2135/tppj2019.03.0006](https://doi.org/10.2135/tppj2019.03.0006).
- Xavier A. 2019. Efficient estimation of marker effects in plant breeding. *G3 (Bethesda).* 9(11):3855–3866. doi:[10.1534/g3.119.400728](https://doi.org/10.1534/g3.119.400728)
- Xavier A, Habier D. 2022. A new approach fits multivariate genomic prediction models efficiently. *Genet Sel Evol.* 54(1):45. doi:[10.1186/s12711-022-00730-w](https://doi.org/10.1186/s12711-022-00730-w).
- Xavier A, Muir WM, Rainey KM. 2020. bWGR: Bayesian whole-genome regression. *Bioinformatics.* 36(6):1957–1959. doi:[10.1093/bioinformatics/btz794](https://doi.org/10.1093/bioinformatics/btz794).
- Xu X, Gao P, Zhu X, Guo W, Ding J, Li C, Zhu M, Wu X. 2019. Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu province, China. *Ecol Indic.* 101:943–953. doi:[10.1016/j.ecolind.2019.01.059](https://doi.org/10.1016/j.ecolind.2019.01.059).
- Yadav D, Jain R, Agrawal H, Chattopadhyay P, Singh T, Jain A, Singh SB, Lee S, Batra D. 2019. EvalAI: towards better evaluation systems for AI agents. arXiv 1902.03570. <https://doi.org/10.48550/arXiv.1902.03570>.
- Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, et al. 2019. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet.* 51(6):1052–1059. doi:[10.1038/s41588-019-0427-6](https://doi.org/10.1038/s41588-019-0427-6).
- Yates F, Cochran WG. 1938. The analysis of groups of experiments. *J Agric Sci.* 28(4):556–580. doi:[10.1017/S0021859600050978](https://doi.org/10.1017/S0021859600050978).
- Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics.* 30(7):1006–1007. doi:[10.1093/bioinformatics/btt730](https://doi.org/10.1093/bioinformatics/btt730).
- Zhou Z-H. 2015. Ensemble learning. In: Li SZ, Jain AK, editors. *Encyclopedia of Biometrics.* Springer US. p. 411–416.

Editor: M. Sillanpää