Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Foundation under Public Law

Stronger Together: New data approach makes plant predictions more accurate

Press Release

Scientific Contact

Prof. Dr. Jochen Reif
Phone: +49 39482 5840
reif@ipk-gatersleben.de

Media Contact
Christian Schafmeister
Phone: +49 39482 5461
schafmeister@ipk-gatersleben.de

Gatersleben, 13.05.2025 **Large amounts of data ("big data") offer enormous potential for improving the accuracy of genome-wide predictions in plant breeding. Encouraged by successful results with wheat hybrids, researchers at the IPK Leibniz Institute have now extended this approach to so-called inbred lines. For the first time, they combined phenotypic and genotypic data from four commercial wheat breeding programmes. The study results were published in the "Plant Biotechnology Journal".**

Deep learning methods have become increasingly crucial in genomic prediction in recent years. In contrast to conventional methods, deep learning approaches work with flexible, non-linear transformations of the input data. The aim is to recognise patterns in the data and link these to observable characteristics such as yield or plant height. The parameters required for this are optimised based on extensive training data. Such methods promise particular advantages when plant characteristics are strongly influenced by complex interactions that are insufficiently considered in conventional models.

In this context, a research team at the IPK has taken on the role of academic data trustee and merged the data from four wheat breeding programmes with trial data from earlier public-private partnerships. "We needed data from many genotypes that had already been tested in different environments, i.e. at different locations", explains Prof. Dr. Jochen Reif, head of the department "Breeding Research" at the IPK.

The new data set covered twelve years of trial activity in 168 environments and formed a training set for genomic predictions with up to 9,500 genotypes - including grain yield, plant height and heading date. One main challenge was merging the different data and ultimately making it comparable. "Despite the heterogeneous phenotypic and genotypic information, we were able to break down the companies' data silos and thus obtain linkable data through meticulous data preparation, including the imputation of missing SNPs", says Prof. Dr. Jochen Reif.

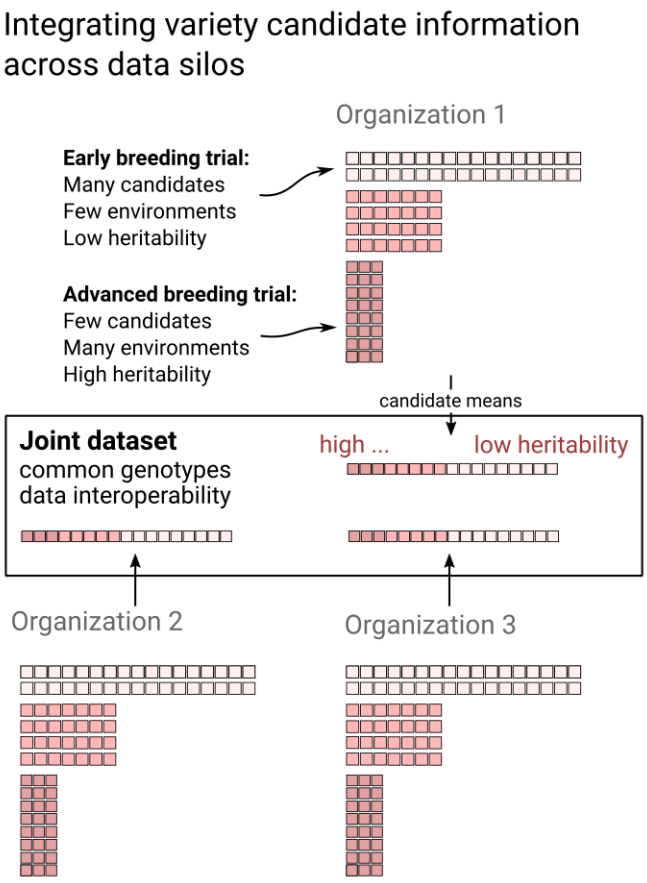
The team used this data to compare classic genomic prediction methods with deep learning approaches based on neural networks. With the help of neural networks, it was possible to recognise patterns in structured data. "Our analyses showed that different test series can be flexibly combined for genomic predictions and that the prediction accuracy continuously improves as the size of the training set increases - at least up to around 4,000 genotypes", explains Moritz Lell, first author of the study. If the training set is increased further, the prediction values increase only slightly.

"However, we assume that this plateau can be overcome if we include significantly more environments in the data set", emphasises Prof. Dr. Jochen Reif. "This would make it possible to utilise the potential of big data in breeding research even better." And this is precisely the aim of the "Drive" project, which has been running since November 2024 and is funded by the Federal Ministry of Education and Research (BMBF).

Original publication:

Lell *et al.* (2025): Breaking down data silos across companies to train genome-wide predictions: A feasibility study in wheat. Plant Biotechnology Journal.
DOI: 10.1111/pbi.70095

Graphic:



By integrating data sources at an academic data trustee more genotypes with high-quality observations can be used for performance prediction.
Graphic: IPK Leibniz Institute/ M. Lell