Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK)
Stiftung des öffentlichen Rechts

Gemeinsam statt einsam: Neuer Datenansatz macht Pflanzenvorhersagen präziser

Pressemitteilung

Wissenschaftlicher Kontakt

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

Medienkontakt
Christian Schafmeister
Tel.: +49 39482 5461
schafmeister@ipk-gatersleben.de

Gatersleben, 13.05.2025 **Große Datenmengen („Big Data“) bieten ein enormes Potenzial, um die Genauigkeit genomweiter Vorhersagen in der Pflanzenzüchtung zu verbessern. Ermutigt durch erfolgreiche Ergebnisse bei Weizenhybriden haben Forschende am IPK Leibniz-Institut diesen Ansatz nun auch auf sogenannte Inzuchtlinien ausgeweitet. Dazu kombinierten sie erstmals phänotypische und genotypische Daten aus insgesamt vier kommerziellen Weizenzuchtprogrammen. Die Ergebnisse der Studie wurden im „Plant Biotechnology Journal“ veröffentlicht.**

In den vergangenen Jahren haben Deep-Learning-Methoden im Bereich der genomischen Vorhersage an Bedeutung gewonnen. Im Gegensatz zu klassischen Verfahren arbeiten Deep-Learning-Ansätze mit flexiblen, nichtlinearen Transformationen der Eingabedaten. Ziel ist es, Muster in den Daten zu erkennen und diese mit beobachtbaren Eigenschaften wie Ertrag oder Pflanzenhöhe zu verknüpfen. Die dafür notwendigen Parameter werden auf der Basis von umfangreichen Trainingsdaten optimiert. Solche Verfahren versprechen insbesondere dann Vorteile, wenn Pflanzeneigenschaften stark von komplexen Wechselwirkungen beeinflusst werden, die in herkömmlichen Modellen nicht oder nur unzureichend berücksichtigt werden.

Ein Forschungsteam am IPK hat in diesem Zusammenhang die Rolle eines akademischen Datentreuhänders übernommen und die Daten aus vier Weizenzüchtungsprogrammen mit Versuchsdaten aus früheren öffentlich-privaten Partnerschaften zusammengeführt. „Wir brauchten im Grunde Daten von vielen Genotypen, die bereits in unterschiedlichen Umwelten, also an unterschiedlichen Standorten, getestet wurden“, erläutert Prof. Dr. Jochen Reif, Leiter der Abteilung „Züchtungsforschung“ am IPK.

Insgesamt umfasste der neue Datensatz zwölf Jahre Versuchstätigkeit in 168 Umwelten und bildete ein Trainingsset für genomische Vorhersagen mit bis zu 9.500 Genotypen - unter anderem zu Kornertrag, Pflanzenhöhe und Ährenschieben. Eine der zentralen Herausforderungen bestand darin, die verschiedenen Daten zusammenzuführen und letztlich vergleichbar zu machen. „Trotz der heterogenen phänotypischen und genotypischen Informationen konnten wir durch eine sehr sorgfältige Datenaufbereitung, inklusive Imputation fehlender SNPs, die Datensilos der Unternehmen aufbrechen und so verknüpfbare Daten gewinnen“, sagt Prof. Dr. Jochen Reif.

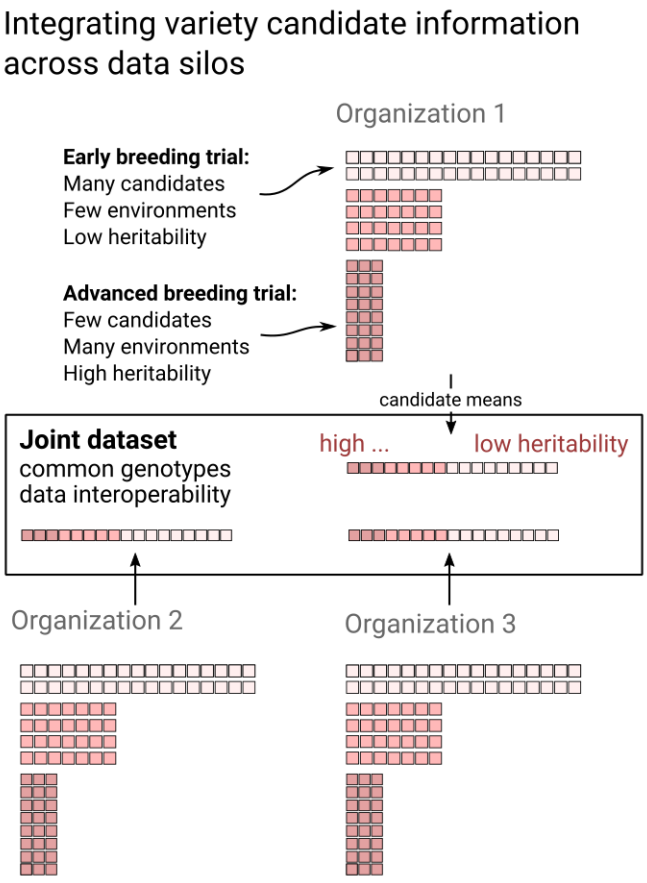
Diese Daten nutzte das Team, um klassische genomische Vorhersagemethoden mit Deep-Learning-Ansätzen auf Basis neuronaler Netzwerke zu vergleichen. Mit Hilfe der neuronalen Netzwerke war es möglich, Muster in strukturierten Daten zu erkennen. „Unsere Analysen zeigten, dass sich verschiedene Versuchsserien flexibel für genomische Vorhersagen kombinieren lassen und sich die Vorhersagegenauigkeit dabei mit wachsender Größe des Trainingssets kontinuierlich verbessert - zumindest bis zu etwa 4.000 Genotypen“, erklärt Moritz Lell, Erstautor der Studie. Wird das Trainingsset darüber hinaus weiter vergrößert, steigen die Vorhersagewerte nur noch geringfügig.

„Wir gehen jedoch davon aus, dass sich dieses Plateau überwinden lässt, wenn wir noch deutlich mehr Umwelten in den Datensatz aufnehmen“, betont Prof. Reif. „Das würde es ermöglichen, das Potenzial von Big Data in der Züchtungsforschung noch besser zu nutzen.“ Und genau das ist auch das Ziel des Projekts „Drive“, das bereits seit November 2024 läuft und vom Bundesministerium für Bildung und Forschung gefördert wird.

Originalpublikation:

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

Grafik:



Durch die Integration von Datenquellen bei einem akademischen Datentreuhänder können mehr Genotypen mit hochwertigen Beobachtungen für die Leistungsvorhersage verwendet werden.
Grafik: IPK Leibniz-Institut/ M. Lell