OXFORD

Briefings in Bioinformatics, 2023, **24(6)**, 1–12
https://doi.org/10.1093/bib/bbad349
Problem Solving Protocol

SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding

Pengfei Gao†, Haonan Zhao†, Zheng Luo, Yifan Lin, Wanjie Feng, Yaling Li, Fanjiang Kong, Xia Li, Chao Fang and Xutong Wang

Corresponding authors: Xia Li, Huazhong Agricultural University, No. 1 Shizishan Road, Hongshan District, Wuhan, Hubei 430070, China. Tel.: +86-027-87282130. Fax: +86-027-87282130. E-mail: xli@mail.hzau.edu.cn; Chao Fang, School of Life Sciences, Guangzhou University, Guangzhou 510006, China. Tel.: +86-020-39366915, Fax: +86-020-39366915. fangchao@gzhu.edu.cn; Xutong Wang, Huazhong Agricultural University, No. 1 Shizishan Road, Hongshan District, Wuhan, Hubei 430070, China. Tel.: +86-027-87282130. Fax: +86-027-87282130. xtwang@mail.hzau.edu.cn

†Pengfei Gao and Haonan Zhao contributed equally to this work.

Abstract

Soybean is a globally significant crop, playing a vital role in human nutrition and agriculture. Its complex genetic structure and wide trait variation, however, pose challenges for breeders and researchers aiming to optimize its yield and quality. Addressing this biological complexity requires innovative and accurate tools for trait prediction. In response to this challenge, we have developed SoyDNGP, a deep learning-based model that offers significant advancements in the field of soybean trait prediction. Compared to existing methods, such as DeepGS and DNNGP, SoyDNGP boasts a distinct advantage due to its minimal increase in parameter volume and superior predictive accuracy. Through rigorous performance comparison, including prediction accuracy and model complexity, SoyDNGP represents improved performance to its counterparts. Furthermore, it effectively predicted complex traits with remarkable precision, demonstrating robust performance across different sample sizes and trait complexities. We also tested the versatility of SoyDNGP across multiple crop species, including cotton, maize, rice and tomato. Our results showed its consistent and comparable performance, emphasizing SoyDNGP's potential as a versatile tool for genomic prediction across a broad range of crops. To enhance its accessibility to users without extensive programming experience, we designed a user-friendly web server, available at http://xtlab.hzau.edu.cn/SoyDNGP. The server provides two features: 'Trait Lookup', offering users the ability to access pre-existing trait predictions for over 500 soybean accessions, and 'Trait Prediction', allowing for the upload of VCF files for trait estimation. By providing a high-performing, accessible tool for trait prediction, SoyDNGP opens up new possibilities in the quest for optimized soybean breeding.

Keywords: soybean; deep learning; genomic selection; trait prediction; web server; crop breeding

INTRODUCTION

Food insecurity is a growing issue, heightened by the increasing world population and the challenges of climate change [1]. Traditional breeding methods, while effective, can be slow and struggle to keep pace with the demands. For instance, they fall short of achieving the annual yield improvement rate of 2.4% needed to double global soybean production by 2050 [2, 3]. To speed up the process, breeders have turned to genomic tools, including genomic selection (GS) [4].

GS is an advanced technique that can make breeding faster and more efficient [5, 6]. It uses genomic prediction models, along with many genetic markers across the genome, to predict how a trait will perform [7–10]. It has been used successfully in both animal and plant breeding, especially in improving traits like crop yield, breeding value, genomic-environs prediction and disease resistance [5, 9, 11–15]. However, using GS effectively relies on many factors, like the size of the training population, heritability of traits, marker density and the prediction model used [16]. Traditional models, such as linear regression models (GBLUP, rrBLUP and Bayesian methods), often struggle with capturing complex non-additive effects [9, 17–19]. This is the context in which deep learning methods, such as DeepGS and DNNGP, can play a role [10, 20]. They use multiple hidden layers to capture complex, non-linear relationships in the data. However, these techniques require large data sets for accurate predictions, which can be a challenge in some cases.

Soybean [*Glycine max* (L). Merr.] is a globally significant crop, providing a rich source of protein and oil for human and animal consumption [21]. In soybean breeding, genomic prediction has already shown its potential [3, 22, 23]. However, there are still hurdles to overcome, like capturing the full range of genetic diversity in soybeans and refining genomic prediction methods.

In this study, we aim to look more closely at how deep learning methods can be used for genomic prediction in soybean breeding. We used a rich source of soybean genomic data, like the genotypes of thousands of soybean samples from the USDA Soybean Germplasm Collection and their phenotypes from the GRIN-global web server [24, 25]. We modeled a deeper neural network

Pengfei Gao is graduate student specializing in Intelligent Agriculture at Huazhong Agricultural University.
Haonan Zhao is graduate student specializing in Intelligent Agriculture at Huazhong Agricultural University.
Zheng Luo is graduate student specializing in Intelligent Agriculture at Huazhong Agricultural University.
Yifan Lin is Research Assistant at Huazhong Agricultural University.
Yaling Li is graduate student specializing in Crop Genetics and Breeding at Huazhong Agricultural University.
Xilong Feng is graduate student specializing in Crop Genetics and Breeding at Huazhong Agricultural University.
Fanjiang Kong is professor at Guangzhou Agricultural University.
Xia Li is professor at Huazhong Agricultural University.
Chao Fang is professor at Guangzhou Agricultural University.
Xutong Wang is professor at Huazhong Agricultural University.
Received: June 22, 2023. **Revised:** September 13, 2023. **Accepted:** September 14, 2023

framework for genomic prediction in soybean called SoyDNGP. We introduce the unique 3D layer input and the convolutional neural network (CNN) architecture of SoyDNGP. We compare SoyDNGP's predictive capabilities with other machine learning (ML) methods and deep learning models like deepGS and DNNGP. Our findings indicate that SoyDNGP consistently outperforms these models, particularly in regression tasks. We test SoyDNGP's applicability across various soybean populations, including wild soybeans, landraces and elite cultivars. Our model demonstrates high predictive accuracy across these diverse populations. We extend the application of SoyDNGP to other species like cotton, maize, rice and tomato. Our model maintains high predictive accuracy, proving its versatility and effectiveness beyond soybeans. To make genomic prediction accessible to a broader audience, we introduce a user-friendly web server for SoyDNGP, featuring a trait lookup and trait prediction tool.

MATERIALS AND METHODS

Data sets used for genomic prediction of soybean

The data used for training and predicting with the SoyDNGP model were procured from two comprehensive online databases: SoyBase and GRIN-Global [25, 26]. The genotype information of a large collection of 20 087 soybean accessions, including 42 509 high-confidence SNPs (single-nucleotide polymorphisms) based on the SoySNP50K iSelect BeadChip, was sourced from SoyBase. The pre-built variant call format (VCF) file, corresponding to version 2 of Williams 82 reference sequences, was utilized. In an attempt to enhance the compatibility of our model with SNP data sets derived from methods other than the 50 K SNP chip, an intersection operation was performed with SNP loci generated from resequencing data. The Beagle 5.4 program (version 22Jul22.46e) was used to phase the SNPs and fill in the missing data [27]. This process resulted in a curated set of 32 032 SNP loci, which were used for model training. However, to maintain a uniform set of annual species and minimize the component of mixed accessions, a selection process was implemented, reducing the number of soybean accessions used for model construction to 13 784. These chosen accessions, part of the USDA Germplasm Collection, are representative of a broad spectrum of landraces and elite cultivars from around the globe.

Phenotypic data for each of these selected soybean accessions were obtained from the GRIN-Global database (https://npgsweb.ars-grin.gov/gringlobal/search). Despite an initial collection of 23 agronomic traits, our focus was narrowed down to 10 key traits. This included six quantitative traits such as protein content (protein), oil content (oil), hundred-seed weight (SdWgt), flowering date (R1), the maturity date (R8), yield and plant height (Hgt). In addition, four qualitative traits were also considered, which encompassed stem termination (ST), flower color (FC), pubescence density (PDENS) and pod color (POD). Information on trait names, along with their corresponding trait ontology (TO) and crop ontology for soybean (CO), is provided in Table S1.

SoyDNGP model structure

In stark contrast to the traditional DNNGP's three-layer wide convolution architecture used for genome-wide big data analysis, our SoyDNGP utilizes a deep and slim network structure [10]. This structure is inspired by the concept of segmentation drawn from the VGG deep learning network [28]. Specifically, SoyDNGP is built around 'convolutional blocks', each incorporating a convolutional layer, a normalization layer and an activation layer (ReLU). The model structure is illustrated in Figure S1.

Every feature extraction unit in the network is comprised of one or two of these convolutional blocks, resulting in an effective block structure for feature extraction. At the end of the convolutional sequence, we have included a fully connected layer to enhance the expression capabilities of the network. With the network's increased depth, we have also added a normalization layer after each convolution to enhance the model's ability to generalize and a dropout layer (dropout = 0.3) to mitigate overfitting. Overall, the network architecture integrates 12 convolutional layers and a single fully connected layer, designed to handle an input tensor of dimensions ($206 \times 206 \times 3$).

The first convolutional module operates using a 3×3 convolution kernel with a stride of one, which effectively upscales features and expands the feature map from three channels to 32. The subsequent convolutional block deploys a 4×4 convolution kernel with a stride of two, increasing the dimensions of the feature map while simultaneously reducing the size of each dimension's feature map. In the network structure that follows, each feature extraction block consists of two convolutional layers.

In each feature extraction block, the first convolutional layer adjusts the convolution kernel size and sampling stride based on the dimensions of the feature map. This guarantees a complete traversal of the feature map while enabling feature map scaling and dimensionality increase with the smallest feasible convolution kernel. The second convolutional layer uses a 3×3 convolution kernel to reprocess the feature map from the preceding layer, reinforcing feature extraction. This process is iterated until the feature map's channel count escalates to 1024, with dimensions reducing to 7×7. Subsequently, the feature map is flattened into a 1D vector and forwarded to the fully connected layer for final classification and regression processing. Given the extensive information density of the SNP variation-based feature matrix, we have chosen to move away from the simplistic zero-padding approach during convolution padding. Instead, we apply a symmetrical filling technique that leverages matrix elements at the outermost layer, using the matrix edge as the axis of symmetry. This significantly bolsters the feature extraction capability from the matrix.

To circumvent the potential issue of overfitting in the model training process induced by the depth of the network, weight decay was applied to the Adam optimizer. This included a decay rate of 1e−5 for regression and 0.01 for classification. For qualitative traits, the model was trained using the commonly applied cross-entropy loss function. Conversely, for regression tasks pertaining to quantitative traits such as protein content and yield, SoyDNGP utilized the smooth L1 loss function ($\beta = 0.1$) as its loss function [Equation (1)]:

$$\text{Smooth L1}: \mathrm{L}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{1}{2\beta}(x_i, y_i)^2, |x_i - y_i| < \beta \\ |x_i - y_i| - \frac{1}{2}\beta, \text{otherwise} \end{cases} \quad (1)$$

This particular loss function provides a constant gradient when the loss is significant, thereby mitigating the potential disruption of training parameters due to substantial gradients. Conversely, when the loss is minimal, the gradient dynamically reduces, addressing the challenge of convergence often seen with L1 loss. Compared to traditional L1 and L2 loss functions, the smooth L1 loss function offers accelerated convergence speed, improved robustness to outliers and enhanced gradient smoothness. For each trait under consideration, we conducted 150 epochs on GeForce RTX 3090 or RTX A6000, selectively preserving the epoch

that demonstrated optimal performance on the test set as the final model weights.

Lastly, it is noteworthy that we have incorporated a coordinate attention (CA) mechanism module after the first and final convolutional layers [29]. This strategy amplifies attention to the positional information in the feature matrix and between channels, thereby enhancing spatial information extraction. SoyDNGP's model structure is designed and implemented using PyTorch (version 2.0.1), a widely recognized open-source ML library [30].

Remodeling deepGS in Python

To facilitate a fair comparison between model architectures, and acknowledging the limited feature representation ability of the original deepGS (rDeepGS) model, we opted to enhance it while maintaining its overall structure [20]. This structure comprises a combination of a convolution layer, a ReLU activation function, a max pooling layer and a dropout layer, all linked to two fully connected layers.

In the rDeepGS model, we substituted the broad 1×18 convolution kernel with a more compact 3×3 kernel. Additionally, we increased the count of both convolution and pooling layers in the model to six, yielding a total of 12 layers, as modified deepGS (mDeepGS). This modification ensured the channel count of the final feature map aligned with that of SoyDNGP. The model structure is illustrated in Figure S2. After adjusting the model structure, we preserved all other conditions identical to those in the SoyDNGP model for the training phase. This approach enabled us to draw an equitable comparison between the two model architectures. Moreover, it underscored the superiority of a slender and deep convolutional network in the realm of feature extraction and representation capabilities.

Model construction of traditional ML algorithms

In order to gage the effectiveness of our proposed SoyDNGP model, we conducted parallel evaluations using nine conventional ML algorithms on identical data sets. These traditional models encompassed: *K*-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Adaptive Boosting (Adaboost), Gaussian Naive Bayes (GNB) and Support Vector Classification (SVC) with different kernels—Linear, Radial Basis Function (RBF) and Sigmoid [31–37]. Each trait was subjected to training using these nine algorithms, facilitating a comparative analysis of their performance and robustness against SoyDNGP on the same data set. The hyperparameter configurations for these models were as follows: In KNN, we assigned the number of neighbors (n_neighbors) as 3. For DT and RF, we confined the maximum depth of the trees (max_depth) to 5, while for RF, we also defined the number of trees in the forest (n_estimators) as 10 and the number of features considered for the optimal split (max_features) as 1. For MLP, we stipulated the L2 penalty (regularization term) parameter (alpha) as 1. The remaining models utilized their default parameters as defined in their respective libraries.

We implemented a 10-fold cross-validation scheme (n_splits = 10) for a more rigorous evaluation of the models, ensuring diverse splits for each run (random_state = None) and random shuffling of the data prior to fold creation (shuffle = True). This was done to preclude the possibility of any class's overrepresentation in any given fold, which might skew the model's performance. Our assessment metrics consisted of precision, recall and F1-score for each class of traits. Additionally, we calculated the mean and SD of accuracy across the folds, offering an encompassing view of the model's performance. To ascertain their generalizability, we evaluated the models based on their accuracy on both training and test data sets.

Data processing of resequencing of soybean database

To assess the performance of SoyDNGP in different soybean populations, we obtained resequencing data from two public data sets available on NCBI (PRJNA608146) and GSA (CRA002269) [38, 39]. The following steps were taken to process all sequencing reads: Initially, Cutadapt (version 3.5.0) was employed to excise potential adaptors and discard low-quality reads [40]. The clean reads were then aligned to version 2 of the Williams 82 reference genome sequences (https://phytozome-next.jgi.doe.gov/) utilizing BWA (version 0.7.17-r1188) [41]. Next, PCR duplicates and reads that mapped to multiple locations were eliminated using SAMtools (version 1.15.1-41-gc7acf84) [41, 42]. The GATK pipeline was subsequently deployed to produce reliable SNPs for variation and evolutionary analysis. SNPs were preserved if they were biallelic and had an MAF greater than 0.05 [43]. To maintain SNP loci that overlapped with the SoyDNGP training data, a total of 32 624 SNPs were chosen using VCFTools (version 0.1.16) [44].

The VCF files generated were employed to predict the phenotypes using pre-built SoyDNGP models. We focused on a population comprising 559 soybean accessions for which three phenotypes, namely, flowering time, hundreds of seeds weight and plant height, had been previously measured. These phenotypes were recorded in 2018 for plants cultivated in Zhengzhou, Henan Province, China (latitude 34.7N, longitude 113.6E). To quantify the accuracy of the phenotype predictions, we computed the Pearson correlation coefficient (r) between the observed phenotypes and the predicted values.

Web server implementation

Our web server is established on a Webflow template (https://webflow.com/), which is enhanced with the Bootstrap5 framework (https://getbootstrap.com/) and operated via the Flask web framework (https://flask.palletsprojects.com/) [45]. To facilitate additional functions, we have chosen several specific tools. Redis (https://redis.io/) functions as the custodian of progress data and prediction results, while MongoDB (https://www.mongodb.com/) is employed to store the data available for users [46, 47]. Gunicorn (https://gunicorn.org/), a Python WSGI HTTP server, manages server operations and Nginx (https://nginx.org/en/) is used for request forwarding from port 80 to the Gunicorn service, as well as for load balancing [48]. The entire project is hosted on a Linux-based system equipped with an i7-13700KF processor and an RTX 3060Ti graphics card. The components of SoyDNGP's Web Server are illustrated in Figure S3.

RESULTS

SoyDNGP exhibits impressive capabilities in soybean genomic prediction

SoyDNGP employs a 3D layer input derived from standard VCF files. In our study, we used data processing libraries such as pandas and numpy to convert VCF files into data matrices. These matrices have sample names as indices and variant sites as column names. Each row of the matrix undergoes resizing to form a 3D matrix of size (M,M,3) [Equation (2)].

$$M = \left\lceil \sqrt{N} \right\rceil \quad N = \text{Number of SNPs} \tag{2}$$

In the input VCF files, there are three types of mutations: 0/0, 0/1, and 1/1. Each type of mutation is represented in a different channel in the feature map, ensuring the relative distances among mutations (Figure 1A). Specifically, 0/1 mutations are indicated in the second channel. The pixel values $p[i,j,k]$ in the feature matrix only have two possible values: 0 and 1. A value of 0 represents the presence of a certain type of mutation at that particular SNP site for the given sample, and a value of 1 indicates the absence of that mutation. Regarding the dimension of our feature matrix, we decided on a size based on the data set with the maximum number of SNP variants, which had 42 000 SNPs. This decision was made to ensure that the model's input would have robustness across different populations. To minimize the impact of missing SNP sites, we repeatedly filled the feature matrix with the sample's own variant features until all pixels were filled. This method allows the SoyDNGP structure to consider both the type of genotype and its spatial relationship. Two distinct structures are deployed for classification (qualitative traits) and regression (quantitative traits) tasks (Figure 1A).

SoyDNGP implements a CNN architecture that is characterized by 12 convolutional layers and one fully connected layer (Figure 1B). During the training phase, the Adam optimizer (Adaptive Moment Estimation), which incorporates principles of momentum and adaptive learning rate methods, is utilized for updating the weights of the model. This optimizer strategy allows for efficient evasion from saddle points and accelerates the model's convergence to optimal fitting. To incorporate attention mechanisms, we compared the performance of coordinate attention (CA), squeeze-and-excitation (SE) and convolutional block attention module (CBAM). Our findings revealed that integrating attention mechanisms substantially enhanced the model's stability and feature representation capabilities (Figure S4). The SE attention mechanism only focuses on channel information [49]. On the other hand, the CABM attention mechanism encompasses both channel and positional information extraction but does not achieve an effective fusion of these features [50]. The CA attention mechanism rectifies these limitations, enabling superior extraction of spatial location information from feature maps [29]. Additionally, with only marginal differences in parameter quantity and floating point operations per second, the CA attention mechanism demonstrates faster fitting speeds during model training. Among the options, CA surpassed SE and CBAM in performance, making it the preferred choice for our final architecture (Figures S4 and S5) [49, 50]. The CA module is strategically placed after the initial and final convolutional layers, enriching the model's ability to focus on both spatial details within the feature matrix and inter-channel correlations (Figure 1C). Subsequently, we experimented with adding more complex residual network modules (Residual Block) to our SoyDNGP model. However, we found that these complex structures increased the number of parameters and computational load without significantly boosting performance (Figure S5). As a result, we opted for a CA + baseline network structure for our final model.

To ascertain the optimal sample size for model training, we trained the model using varying numbers of samples and monitored the predictive performance. The samples were divided into groups of 2 k, 5 k, 8 k and 10 k for training, each paired with test sets of 11 784, 8784, 5784 and 3784 samples, respectively, over a span of 150 epochs. Our findings indicated that a sample size of 2 k yields lower performance in terms of accuracy and other metrics, while no significant differences were observed among larger sample sizes (Figures S6 and S7). Ultimately, we found a sample size of 5 k to be the most suitable for model construction.

We then conducted individual predictions to test accuracy. The results revealed that the prediction accuracy for regression tasks ranged from 0.56 in R8 to 0.87 in SdWgt, while for classification tasks, it ranged from 0.82 in ST to 0.96 in FC (Figure 1D). This conclusion was also supported by the absolute errors between normalized observed and predicted phenotypic values, as depicted in Figure S8. Through extensive testing, the model consistently delivers impressive prediction accuracy in both regression and classification tasks. In our study, some traits indeed exhibit imbalanced class distributions, leading to poorer model performance in underrepresented categories. However, the model performs exceptionally well for phenotypes with relatively balanced class distributions (Figure S9). For instance, in the case of 'Flower color', which has a balanced binary classification, the model performs well. In contrast, for 'H_CLR,' the model's accuracy for 'Br' and 'Bl' classes is noticeably lower, a result we attribute to the skewed distribution of these classes in the data set (Figure S9).

Comparative performance of SoyDNGP and other algorithms in trait prediction

In order to evaluate the performance of SoyDNGP in genomic prediction relative to other ML methodologies, we utilized an identical data set for training SoyDNGP models, which was also applied to other procedures. Although conventional ML techniques are not optimized for regression tasks, we discovered that several were capable of performing classification tasks with high accuracy. For instance, the DT model yielded prediction accuracies of 0.97 and 0.85 for FC and POD, respectively (Figure 2A). With the SVMRBF model, the accuracy of ST and PDENS reached 0.82 and 0.84, respectively. Among the nine ML methods tested, SoyDNGP exhibited balanced performance across all classification traits, with accuracies ranging from 0.82 (ST) to 0.94 (FC) (Figure 2A).

To assess the performance of SoyDNGP in comparison with other CNN-based deep learning models, such as deepGS and DNNGP, we recreated their model architectures according to the details provided in the original research literature using Python. Regrettably, the original version of deepGS (rDeepGS) performed subpar and was unsuitable for regression tasks, despite its comparable performance in classification tasks with other methods (Figure 2B and Table 1). To confirm the efficiency of the deepGS structure, we re-engineered it into a modified version (mDeepGS). Training these models with the same data set as used for SoyDNGP revealed that regardless of the trait or the amount of training samples utilized in regression tasks, SoyDNGP showed improved performance compared to both mDeepGS and DNNGP in our tests (Figure 2B). The correlation coefficient (r) for DNNGP deviated by approximately 5% from that of SoyDNGP (Figure 2B, upper panel). Moreover, the discrepancy between predicted and actual values (measured by mean squared error, MSE) was nearly 10 times larger than with SoyDNGP (Figure 2B, lower panel). This suggests that DNNGP only has the capacity to predict trends and qualitatively describe them but also lacks precision in quantification. Owing to its shallow structure, mDeepGS was unable to effectively manage the complexity of the regression task, thus failing to fit accurately. Our observations revealed that the three deep learning models—DeepGS, DNNGP and SoyDNGP—showed comparable performances for qualitative trait classification tasks. However, significant differences were observed in their performances on regression tasks. rDeepGS, much like traditional ML models, failed to effectively fit the regression tasks. This could

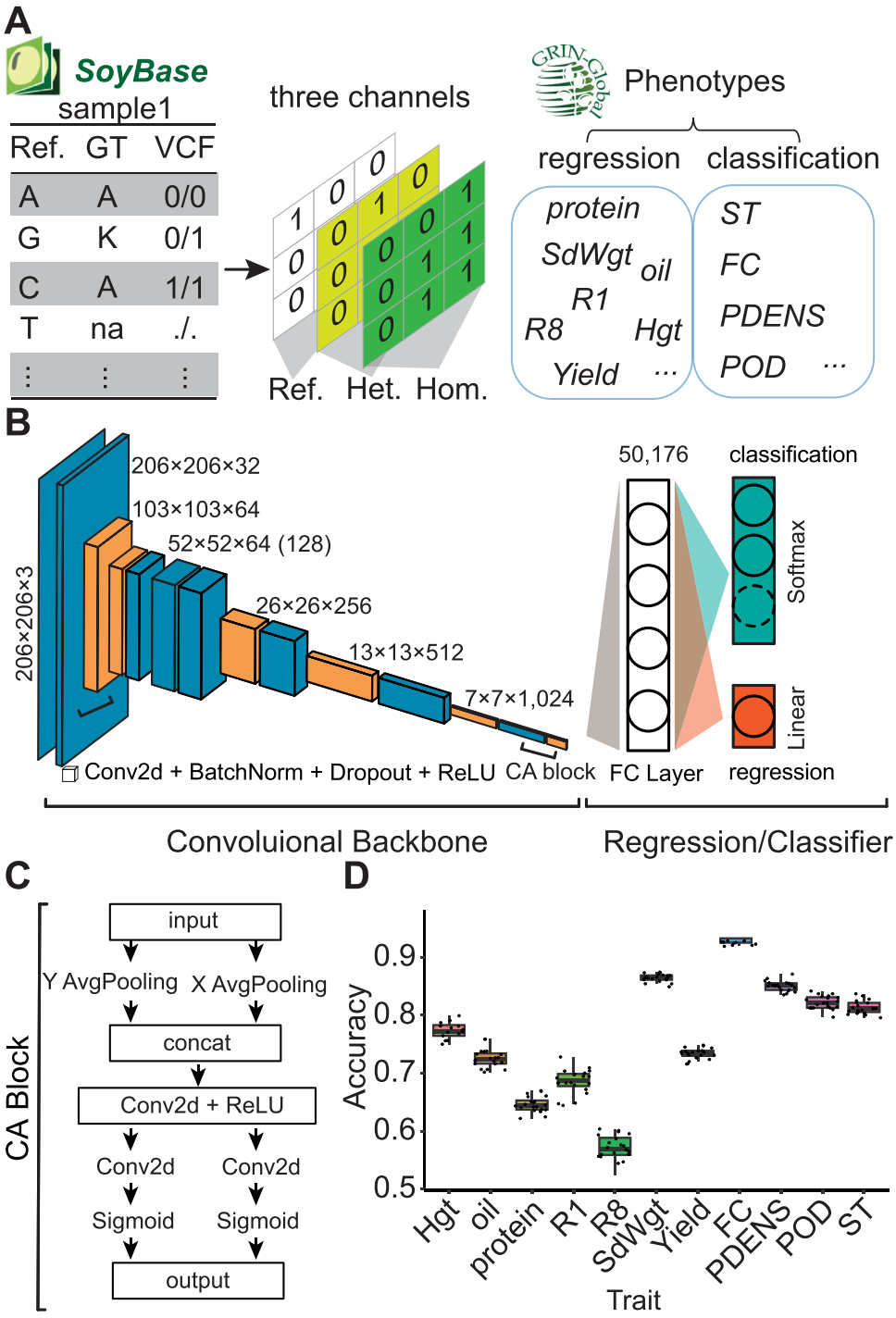

Figure 1. Overview of SoyDNGP's features. (**A**) The transformation process of genotype and phenotype data as input for SoyDNGP. (**B**) Depiction of the SoyDNGP module structure for classification and regression tasks. (**C**) Detailed illustration of the CA Block. (**D**) SoyDNGP's predictive accuracy for 11 key agronomic traits.

be primarily attributed to the less complex nature of classification tasks that can be effectively tackled with ML techniques, resulting in minimal differences in model performance across these tasks. Additionally, we found that rDeepGS and mDeepGS have shorter run times but their performance was unsatisfactory. SoyDNGP and DNNGP had nearly identical run times, yet SoyDNGP had over 10 times the parameter volume of DNNGP. This higher parameter volume allowed SoyDNGP to better learn and fit more complex features, demonstrating stronger generalization capabilities (Table S2). This evidence indicates that the SoyDNGP model structure holds a significant advantage in genomic prediction compared to other methods.

Versatile predictive capacity of SoyDNGP across diverse soybean populations

Our model was developed using the USDA soybean germplasm collections, leaving us uncertain about its application to other resources across diverse countries and latitudes. To appraise the predictive prowess of our constructed models via SoyDNGP, we applied it to a soybean population comprising 559 accessions,

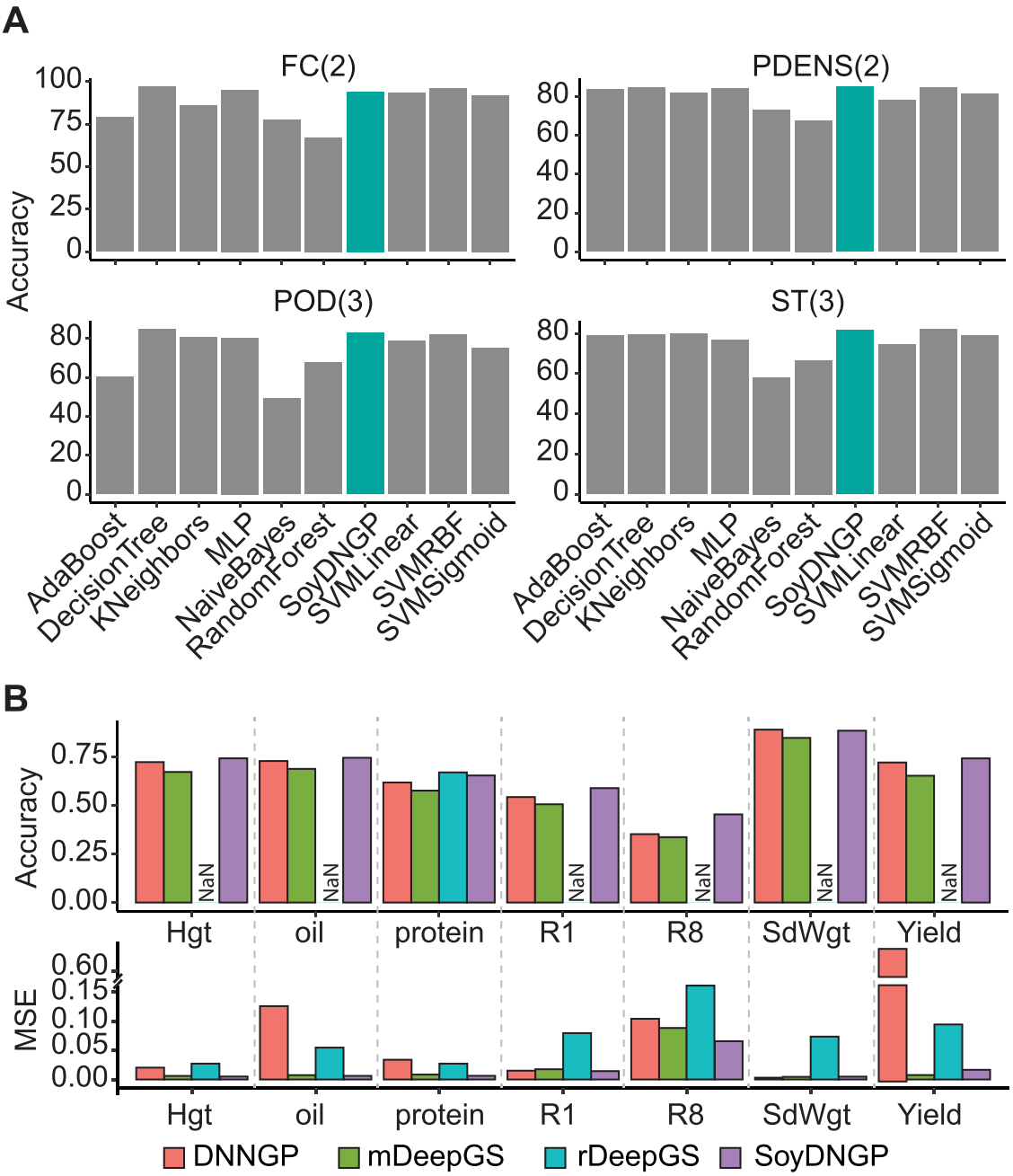

Figure 2. Comparative analysis of predictive performance between SoyDNGP and other approaches. (**A**) The predictive accuracy of SoyDNGP in comparison with traditional machine learning methods for classification tasks such as FC, PDENS, POD and ST. The numbers in parentheses on the plot denote the number of classification categories. (**B**) The predictive accuracy of SoyDNGP in comparison with other deep learning-based methods for regression tasks. Hgt, Oil, Protein, R1, R8, SdWgt and Yield represent plant height, oil content, protein content, flowering time, maturity time, hundred seed weight and yield, respectively. Accuracy is quantified by the correlation coefficient (*r*). 'MSE' denotes mean squared error, reflecting the absolute errors between the normalized observed and predicted phenotypic values.

inclusive of 121 wild soybeans (*G. soja*), 207 landraces and 231 elite cultivars [38]. We executed predictions for 16 qualitative traits and 12 quantitative traits (Table S3). To substantiate the prediction accuracy for significant yield and quality traits, we juxtaposed the phenotypes of specified soybean traits grown in Zhengzhou, China in 2018 with our predictions. Our analysis unveiled a robust positive correlation between predicted and actual values (Figure 3A). For instance, the correlation for R1 and Hgt stood at 0.56 and 0.51, respectively. Most impressively, the prediction accuracy for the SdWgt reached an exceptional 0.84 (Figure 3A). These results suggest that our prediction models bear wide applicability across diverse soybean populations. One potential explanation for the high prediction accuracy for seed weight across different populations could be that environmental factors play a lesser role in this trait compared to others, such as R1 and Hgt (Figure 3A).

Remarkably, even though wild soybean was not included in model training, our model remains useful for predicting the traits of wild soybean (Figures 3B and S10). For example, our predictions indicated high protein content and lower oil content and yield for wild soybean compared to landrace and elite cultivars (Figure 3B), a finding consistent with prior soybean research [51, 52]. This

Table 1. The predictive accuracy of SoyDNGP in comparison with other deep learning-based methods for classification tasks

Trait	Methods	Accuracy	Precision	Recall	f1 score
FC	rDeepGS	0.95	0.94	0.94	0.94
	mDeepGS	0.67	0.34	0.50	0.40
	DNNGP	0.94	0.93	0.94	0.94
	SoyDNGP	0.94	0.93	0.93	0.93
PDENS	rDeepGS	0.85	0.84	0.83	0.83
	mDeepGS	0.61	0.30	0.50	0.38
	DNNGP	0.84	0.83	0.83	0.83
	SoyDNGP	0.85	0.84	0.85	0.84
POD	rDeepGS	0.82	0.77	0.69	0.71
	mDeepGS	0.68	0.23	0.34	0.27
	DNNGP	0.80	0.76	0.66	0.69
	SoyDNGP	0.83	0.80	0.70	0.73
ST	rDeepGS	0.81	0.71	0.63	0.64
	mDeepGS	0.54	0.18	0.33	0.23
	DNNGP	0.80	0.67	0.61	0.61
	SoyDNGP	0.82	0.68	0.63	0.63

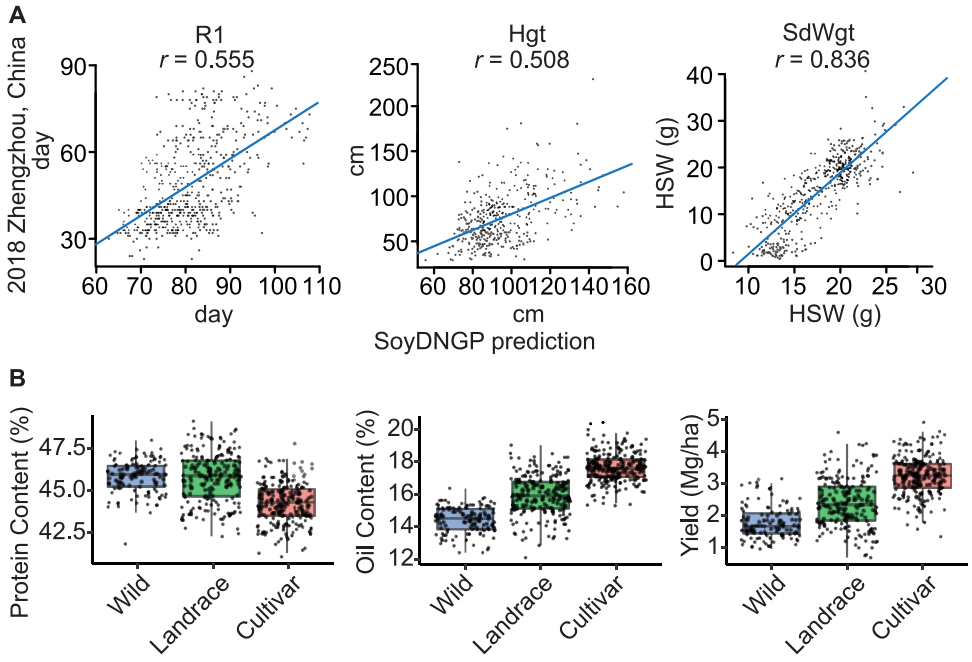

Figure 3. Evaluation of SoyDNGP's predictive capacity in diverse soybean populations. (**A**) Comparison of observed and predicted phenotypes for selected soybean traits cultivated in Zhengzhou, China in 2018 using the SoyDNGP model. The trend line depicts linear regression. (**B**) Distribution of predicted phenotypes for a given trait across three distinct subpopulations.

also implies that gene exchange between wild and domesticated soybean might be facilitated by significant gene flow [53].

Expansive application of SoyDNGP beyond soybean

In an effort to evaluate the versatility and efficacy of SoyDNGP, we put it to test with other species, using genotype data and five representative traits from cotton, maize, rice and tomato populations [54–57]. For the sake of comparison, the same data sets were also applied to DNNGP and mDeepGS. Apart from mDeepGS, which exhibited the lowest accuracy, SoyDNGP demonstrated predictive accuracies ranging from an average of 0.50 in maize to an average of 0.71 in rice (Figure 4A). A similar performance spectrum was observed in DNNGP (0.49–0.69) (Figure 4A). It's noteworthy that for smaller sample sizes such as maize and tomato, with 214 and 508 samples, respectively, DNNGP outperformed SoyDNGP (Figure 4A). However, in larger sample populations like cotton and rice, exceeding 1000 samples, SoyDNGP proved superior (Figure 4A). Despite the similarities in accuracy, DNNGP's mean squared error (MSE) was generally higher than that of SoyDNGP (Figure 4B). Based on these findings, we can conclude that SoyDNGP not only is capable of training and predicting phenotypes of traits in other species but also surpasses the performance of other methods, thereby confirming its robust versatility and effectiveness. Therefore, SoyDNGP stands as a promising tool for genomic prediction, with its application potentially extending beyond soybeans to other crops and organisms, thereby bolstering the advancements in genomics and breeding research.

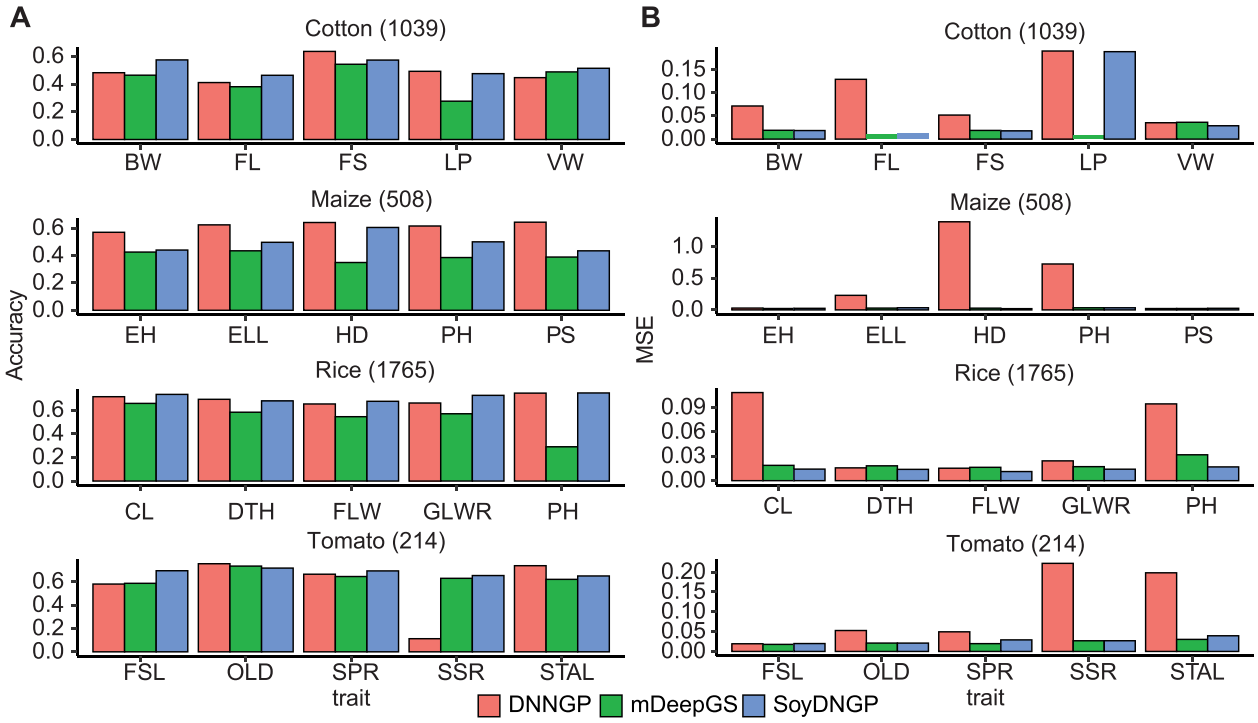

Figure 4. Comparative evaluation of predictive capacity among SoyDNGP and two other methods across diverse crop species. (**A**) Predictive accuracy of various traits assessed by different methods, quantified by the correlation coefficient (*r*). (**B**) Mean standard error (MSE) between normalized observed and predicted phenotypic values for specific traits using different methods. The number in parentheses represents the sample size for a given population. Traits in cotton include boll weight (BW), fiber strength (FS), fiber length (FL) and verticillium wilt (VW). Traits in maize include ear height (EH), ear leaf length (ELL), heading date (HD), plant height (PH) and pollen shed (PS). Traits in rice include culm length (CL), days to heading (DTH), flag leaf width (FLW), PH and grain length–width ratio (GLWR). Traits in tomato include culm length (FSL), days to heading (OLD), sepal length to petal length ratio (SPR), stamen length to (stigma length + ovary longitudinal diameter) ratio (SSR) and stamen length (STAL).

SoyDNGP is an open-friendly web server for the genomic prediction of soybean

To make SoyDNGP accessible to users without deep programming expertise, we have built a web server that bears the same name as our model structure and is available at http://xtlab.hzau.edu.cn/SoyDNGP. The SoyDNGP platform provides two easy-to-use interfaces for exploring trait information. The first feature, 'Trait Lookup', lets users enter the taxon identifier, which could be the plant introduction (PI) number or traditional name, to check whether the corresponding record is already in our database (Figure 5A). Additionally, our 'Trait Lookup' section includes pre-existing trait predictions for 500 soybean accessions, which are in addition to the ones from the USDA soybean germplasm collection, and all have available re-sequencing data [39]. We are continuously increasing this number with daily updates to offer an ever-expanding data set to our users. This functionality can be beneficial for users wishing to select specific soybean accessions based on certain trait predictions, thereby enhancing the efficiency of SoyDNGP. The second feature, the 'Trait Prediction' tool, allows users to upload a VCF file, which our robust predictive models then use to predict trait values (Figure 5A and B). We also provide users with the option to contribute to the enrichment of our lookup database. If users opt to contribute, they will not need to rerun the prediction when revisiting their results in the future.

DISCUSSION

Challenges in existing models

While deepGS and DNNGP have made notable contributions to the field, they present certain limitations that necessitate further research [10, 20]. Specifically, the use of 1D vectors for model input in both deepGS and DNNGP can be limiting when representing complex SNP locus feature information. This simplistic approach may not capture the full depth of genotypic variations, thereby affecting the model's predictive accuracy. Additionally, the shallow, wide convolution architecture used in these models may not be optimal for capturing intricate relationships within the data. Given the limitations in earlier works, there was a clear research gap in developing a model that not only improves predictive accuracy but also addresses the shortcomings in data representation and computational efficiency.

Features of SoyDNGP: addressing the gaps

This study aims to fill these gaps by introducing SoyDNGP, a model that employs a more complex 3D matrix for input features and employs a more rational structure for data processing. SoyDNGP has several significant advantages over deepGS and DNNGP models: enhanced feature density, optimized structure, minimized feature loss, stable training through regularization and incorporation of attention. Unlike deepGS and DNNGP, which use 1D SNP vectors as inputs, SoyDNGP uses a 3D matrix that includes positional and mutational information. This makes it more suitable for CNNs and offers richer feature density. While DNNGP and deepGS use shallow and wide convolutional layers, SoyDNGP employs a deeper and narrower architecture with stacked small kernels for better feature extraction and efficiency. Instead of using max-pooling like deepGS, SoyDNGP uses a convolutional stride of two, effectively fusing and downsampling features with minimal loss [20]. SoyDNGP integrates Dropout and Batchnorm between convolutions and employs L2 regularization, enhancing

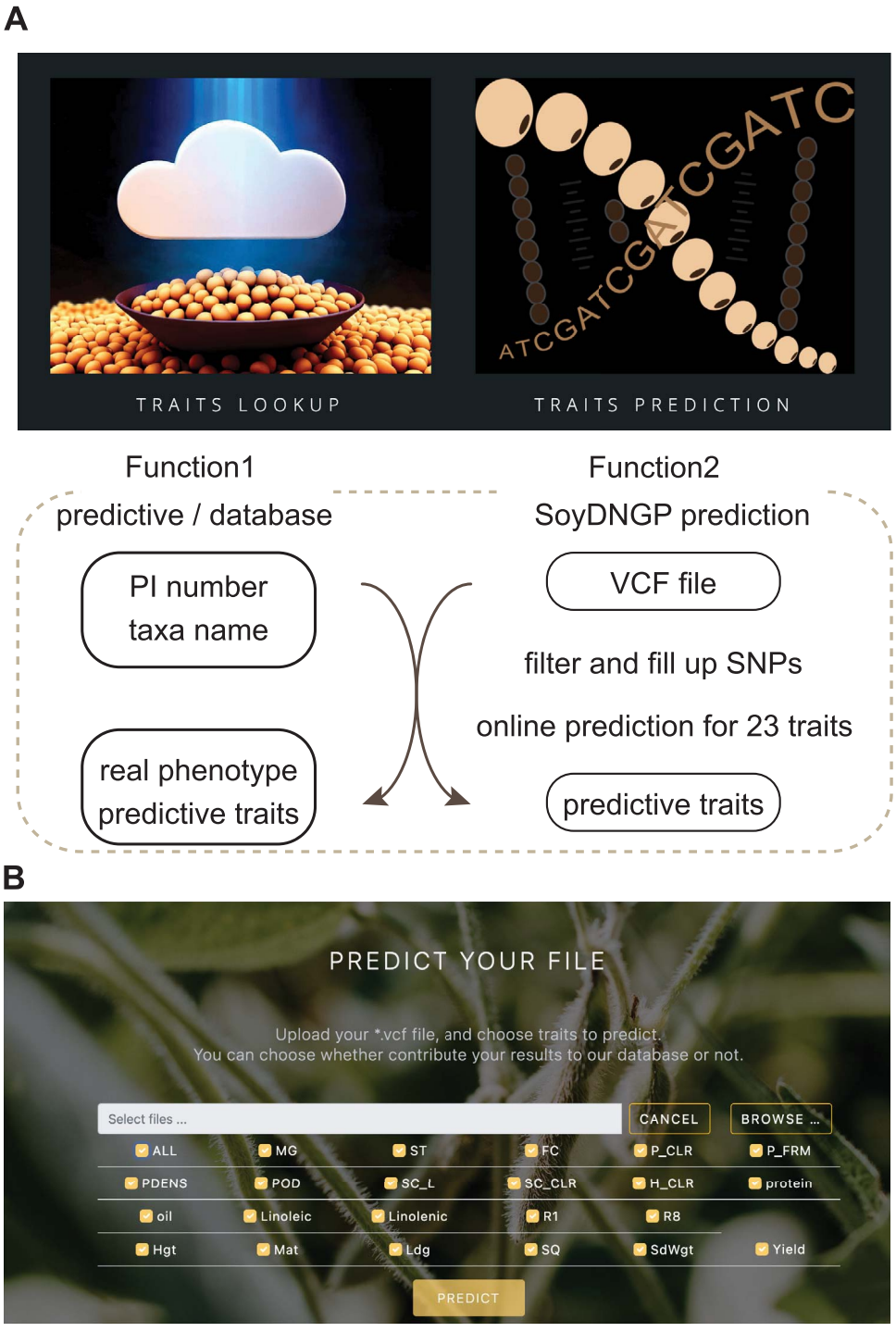

Figure 5. The user interface and functionalities of the SoyDNGP web server. (**A**) Descriptions of the two main functions built into the web server are provided. The upper panel showcases the entrance interface on the webpage, while the lower panel outlines the workflow of the two functions. (**B**) Display of the SoyDNGP prediction function page within the web server. The platform allows users to upload their own VCF files to generate predictions for 28 traits.

model stability and preventing overfitting more effectively than its predecessors [58]. SoyDNGP uses CA mechanisms to consider spatial and channel information, thereby improving its feature extraction capabilities.

Data challenges and the rising need for deeper models

However, our study faces two primary challenges with the data set. The first is an imbalanced sample distribution. Many traits under consideration have multiple categories, often with intricate subdivisions, which lead to a skewed distribution of sample numbers across these classes. This imbalance poses a challenge in training a robust model effectively. The second challenge is data reliability. Traits such as plant height, flowering time and maturity time are often measured without standardized protocols, resulting in significant errors during data collection that affect the model's predictive performance. Our experiments show that as the sample size increases, shallow neural networks like DNNGP

and DeepGS begin to lose their efficacy in quantitatively representing traits. Given the fast-paced advancements in biotechnology, the need for deeper models like SoyDNGP is increasingly becoming evident. Our focus remains on model interpretability in the GS domain, as it is more critical here than in other computational disciplines like image recognition or natural language processing. We have designed our model to be as interpretable as possible, minimizing irreversible operations like pooling. This aligns with our broader goal of identifying crucial gene locations possibly correlated with different traits.

Toward a universal platform in GS

Moreover, there's a lack of a universally adaptable deep learning platform in the GS domain, similar to what YOLO or BIOBERT provides in their respective fields [59, 60]. While Kumar *et al.* [61] recently introduced the DeepMap, it has limitations in terms of flexibility and scalability. To address this, we have developed the *SoyDNGPNext* PyPI package. Based on the baseline SoyDNGP algorithm, this package allows users to easily reconstruct models, train data and make predictions through simple Python commands, thereby enhancing the model's adaptability to various data sets.

In conclusion, we have created and validated SoyDNGP, a CNN-based model tailored specifically for predicting soybean traits. The results underscored that SoyDNGP consistently superseded deepGS and DNNGP models, exhibiting higher accuracy with reduced model complexity. Moreover, we tested SoyDNGP's adaptability across an array of crop species, including cotton, maize, rice and tomato, highlighting its potential as a resilient and versatile tool for genomic prediction. To expand SoyDNGP's reach, we established a user-friendly web server that offers users easy access to trait predictions and the ability to calculate traits using VCF files. Moving forward, our efforts will concentrate on consistently augmenting the database of pre-existing trait predictions and enhancing the accuracy and efficiency of the model. With the model packaged as an accessible PyPI program and integrated into a user-centric web-server—the first of its kind for soybean trait prediction—breeders, even those without bioinformatics backgrounds, can easily predict traits from genotype data sets. This is especially crucial in scenarios where immediate phenotype knowledge is paramount for breeding or disease research. Our model's accuracy aids in streamlining the selection process, enabling researchers to swiftly pinpoint promising plants or progeny. Moreover, our model's precision is the foundation for identifying pivotal genetic features contributing to specific traits, marking our forward direction.

Key Points

- We developed SoyDNGP, a convolutional neural network-based model that significantly advances soybean trait prediction by outperforming existing methods in both parameter volume and predictive accuracy.
- SoyDNGP demonstrated remarkable precision in predicting complex traits across different sample sizes and trait complexities, showcasing its robustness and versatility for use in other crop species.
- A user-friendly web server has been designed for SoyDNGP, enabling researchers with varied programming experience to access and use the model effectively, thereby opening up new possibilities for efficient soybean breeding.
- SoyDNGP's versatility was validated across multiple crop species, establishing its utility as a broad tool for genomic prediction, not limited to soybean.
- The combination of superior performance and accessibility makes SoyDNGP a potential game-changer in the field of crop breeding and genomic analysis.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordbjournals.org/.

ACKNOWLEDGEMENTS

We express our gratitude to Dr Ting Zhao of Zhejiang University for the provision of cotton population genotype and phenotype data; we are also thankful to Dr Xin Wang and Dr Ning Yang from Huazhong Agricultural University for their contribution of tomato and maize data sets.

FUNDING

The National Key Research and Development Program of China (grant 2022YFD1201502).

DATA AVAILABILITY

Complete data sets can be found within the main text, supplementary materials and referenced studies, as well as in public databases. The code, pre-built models and the standalone network structure are accessible at https://github.com/IndigoFloyd/SoybeanWebsite and https://github.com/IndigoFloyd/SoybeanWebsite. In addition to this, the source code detailed in our manuscript has been deposited in Figshare and can be accessed using the following DOI: https://doi.org/10.6084/m9.figshare.23537067.v2. We have packaged our code and uploaded it to PyPi for easier access. The package can be installed by running 'pip install SoyDNGPNext' in the terminal.

REFERENCES

1. FAO, IFAD, UNICEF, WFP and WHO. *The State of Food Security and Nutrition in the World 2021. Transforming Food Systems for Food Security, Improved Nutrition and Affordable Healthy Diets for All*. Rome: FAO, 2021.
2. Ray DK, Ramankutty N, Mueller ND, *et al.* Recent patterns of crop yield growth and stagnation. *Nat Commun* 2012;**3**:1293.
3. Yoosefzadeh-Najafabadi M, Rajcan I, Eskandari M. Optimizing genomic selection in soybean: an important improvement in agricultural genomics. *Heliyon* 2022;**8**:e11873.
4. Decker JE. Agricultural genomics: commercial applications bring increased basic research power. *PLoS Genet* 2015;**11**:e1005621.
5. Bhat JA, Ali S, Salgotra RK, *et al.* Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front Genet* 2016;**7**:221.

6. Bali N, Singla A. Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. *Arch Comput Methods Eng* 2022;**29**:95–112.
7. Sandhu K, Patil SS, Pumphrey M, *et al.* Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. *Plant Genome* 2021;**14**:e20119.
8. Hayes B, Goddard M. Genomic selection. *J Animal Breed Genet* 2007;**8**:323.
9. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29.
10. Wang K, Abid MA, Rasheed A, *et al.* DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol Plant* 2023;**16**:279–93.
11. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 2014;**19**:592–601.
12. Poland J, Rutkoski J. Advances and challenges in genomic selection for disease resistance. *Annu Rev Phytopathol* 2016;**54**:79–98.
13. Shahsavari M, Mohammadi V, Alizadeh B, *et al.* Application of machine learning algorithms and feature selection in rapeseed (*Brassica napus* L.) breeding for seed yield. *Plant Methods* 2023;**19**:57.
14. Newman SJ, Furbank RT. Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. *Nat Plants* 2021;**7**:1354–63.
15. Xu Y, Zhang X, Li H, *et al.* Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Mol Plant* 2022;**15**:1664–95.
16. Xu Y, Crouch JH. Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci* 2008;**48**:391–407.
17. Van Raden PM. Efficient methods to compute genomic predictions. *J Dairy Sci* 2008;**91**:4414–23.
18. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 2011;**4**:250–255.
19. De Los CG, Naya H, Gianola D, *et al.* Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 2009;**182**:375–85.
20. Ma W, Qiu Z, Song J. *et al.* A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 2018;**248**:1307–1318.
21. Hartman GL, West ED, Herman TK. Crops that feed the world 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Secur* 2011;**3**:5–17.
22. Ravelombola W, Qin J, Shi A, *et al.* Genome-wide association study and genomic selection for yield and related traits in soybean. *PloS One* 2021;**16**:e0255761.
23. Stewart-Brown BB, Song Q, Vaughn JN, *et al.* Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3* 2019;**9**:2253–65.
24. Song Q, Hyten DL, Jia G, *et al.* Fingerprinting soybean germplasm and its utility in genomic research. *G3* 2015;**5**:1999–2006.
25. Postman J, Hummer K, Ayala-Silva T *et al.* GRIN-Global: an international project to develop a global plant genebank information management system. *Acta Hortic.* 2010;**859**:49–55.
26. Grant D, Nelson RT, Cannon SB, *et al.* SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res* 2010;**38**:D843–6.
27. Ayres DL, Darling A, Zwickl DJ, *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol* 2012;**61**:170–3.
28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
29. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*. Nashville, TN, USA, 2021, pp. 13713–22.
30. Imambi S, Prakash KB, Kanagachidambaresan G. *PyTorch, Programming with TensorFlow: Solution for Edge Computing Applications*, Springer, Cham, 2021, 87–104.
31. Peterson LE. K-nearest neighbor. *Scholarpedia* 2009;**4**:1883.
32. Myles AJ, Feudale RN, Liu Y, *et al.* An introduction to decision tree modeling. *J Chemom* 2004;**18**:275–85.
33. Biau G, Scornet E. A random forest guided tour. *Test* 2016;**25**: 197–227.
34. Ramchoun H, Ghanou Y, Ettaouil M, *et al.* Multilayer perceptron: architecture optimization and training. *IJIMAI* 2016;**4**:26–30.
35. Feng D-C, Liu Z-T, Wang X-D, *et al.* Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach. *Construct Build Mater* 2020;**230**:117000.
36. Ontivero-Ortega M, Lage-Castellanos A, Valente G, *et al.* Fast Gaussian Naïve Bayes for searchlight classification analysis. *Neuroimage* 2017;**163**:471–9.
37. Hsu C-W, Chang C-C, Lin C-J. *A Practical Guide to Support Vector Classification*. Department of Computer Science, National Taiwan University, Taipei, 2003, 1396–400.
38. Lu S, Dong L, Fang C, *et al.* Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. *Nat Genet* 2020;**52**:428–36.
39. Liu Y, Du H, Li P, *et al.* Pan-genome of wild and cultivated soybeans. *Cell* 2020;**182**:162–176. e113.
40. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10–2.
41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. 2013.
42. Danecek P, Bonfield JK, Liddle J, *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:giab008.
43. McKenna A, Hanna M, Banks E, *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
44. Danecek P, Auton A, Abecasis G, *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.
45. Hammer MM, Kotecha N, Irish JM, *et al.* WebFlow: a software package for high-throughput analysis of flow cytometry data. *Assay Drug Dev Technol* 2009;**7**:44–55.
46. Gade AN, Larsen TS, Nissen SB, *et al.* REDIS: a value-based decision support tool for renovation of building portfolios. *Build Environ* 2018;**142**:107–18.
47. Banker K, Garrett D, Bakkum P, *et al. MongoDB in Action: Covers MongoDB Version 3.0*. Simon and Schuster, Manning Publications Co., New York, 2016.
48. Reese W. Nginx: the high-performance web server and reverse proxy. *Linux J* 2008;**2008**:2.
49. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020;**42**:2011–202.
50. Woo S, Park J, Lee J-Y *et al.* Cbam: Convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. (Eds.), *Computer Vision – ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Volume 11211 of Lecture Notes in Computer Science*, Springer, 2018, pp. 3–19.

51. Swarm SA, Sun L, Wang X, *et al.* Genetic dissection of domestication-related traits in soybean through genotyping-by-sequencing of two interspecific mapping populations. *Theor Appl Genet* 2019;**132**:1195–209.
52. Zhang D, Sun L, Li S, *et al.* Elevation of soybean seed oil content through selection for seed coat shininess. *Nat Plants* 2018;**4**:30–5.
53. Wang X, Chen L, Ma J. Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. *Genome Biol* 2019;**20**:1–15.
54. Wang W, Mauleon R, Hu Z, *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018;**557**:43–9.
55. Ye J, Wang X, Wang W, *et al.* Genome-wide association study reveals the genetic architecture of 27 agronomic traits in tomato. *Plant Physiol* 2021;**186**:2078–92.
56. Liu H, Luo X, Niu L, *et al.* Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. *Mol Plant* 2017;**10**:414–26.
57. Ma Z, He S, Wang X, *et al.* Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat Genet* 2018;**50**:803–13.
58. Cortes C, Mohri M, Rostamizadeh A. L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653. 2012.
59. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.
60. Lou H, Duan X, Guo J, *et al.* DC-YOLOv8: small-size object detection algorithm based on camera sensor. *Electronics* 2023;**12**:2323.
61. Kumar A, Sundaram KT, Gnanapragasam N, *et al.* DeepMap: a deep learning-based model with four-line code for prediction-based breeding in crops. bioRxiv, 2023: 2023.07. 26.550275.