agronomy

MDPI

Article

Explainable TabPFN-Based Machine Learning for Single-Plant Yield Estimation and Trait Prioritization in Faba Bean (*Vicia faba* L.)

Yeter Çilesiz [1], İlkay Yelmen [2], Tolga Karaköy [3], Halit Bakır [4], Seda Karateke [5,*] and Metin Zontul [4]

[1] Department of Field Crops, Faculty of Agricultural Sciences and Technology, Sivas University of Science and Technology, Sivas 58000, Türkiye
[2] Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istinye University, Istanbul 34396, Türkiye
[3] Department of Plant Protection, Faculty of Agricultural Sciences and Technologies, Sivas University of Science and Technology, Sivas 58000, Türkiye
[4] Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Sivas University of Science and Technology, Sivas 58000, Türkiye
[5] Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istanbul Atlas University, Istanbul 34408, Türkiye
* Correspondence: seda.karateke@atlas.edu.tr

Abstract

Faba bean yield reflects complex relationships among genotype, environment, and agronomic traits. This study evaluated an explainable Tabular Prior-data Fitted Network (TabPFN) framework for estimating plot-mean single-plant yield and prioritizing traits using 398 plot-level observations, 13 measured agronomic predictors, and six derived features. On the reference 80/20 split, TabPFN achieved the best values for all four test metrics (R^2 = 0.8746, RMSE = 1.9132 g plant^{-1}, MAE = 1.0819 g plant^{-1}, and MAPE = 8.16%). The Friedman test detected differences among the six models ($\chi^2(5) = 16.75$, $p = 0.005$); Nemenyi comparisons distinguished TabPFN from HistGradientBoosting and SVR, whereas the Holm-corrected Wilcoxon analysis confirmed only the TabPFN–SVR difference. Across 10 repeated 80/20 splits, TabPFN obtained the highest mean test R^2 (0.8614 ± 0.0691), ranked first in eight splits, and produced a higher R^2 than every tuned baseline in at least eight splits. SHAP, permutation importance, and LOCO analyses emphasized pod-, seed-, and biomass-related predictors. Repeated-split ablation showed that derived features improved TabPFN consistently, whereas removing selected target-proximal yield variables reduced performance for every model. The framework is therefore a harvest-time trait-estimation and trait-prioritization tool rather than an early-season forecasting system. Notably, TabPFN achieved this performance without the 100-trial Optuna search used for each baseline; only n_estimators was screened over four prespecified values.

Keywords: *Vicia faba* L.; explainable artificial intelligence (XAI); trait-based yield estimation; agronomic traits; TabPFN; SHAP

Academic Editor: Paul Kwan

Received: 22 June 2026
Revised: 8 August 2026
Accepted: 19 August 2026
Published: 28 August 2026

1. Introduction

Faba bean (*Vicia faba* L.) is one of the most important grain legume crops widely cultivated worldwide due to its high protein content, its significance in human and animal nutrition, its biological nitrogen fixation capacity, and its contributions to sustainable agricultural systems [1]. Particularly cultivated for centuries in the Mediterranean basin, North Africa, Europe, and West Asia, the faba bean is considered a strategically important field crop because of its adaptability to diverse agroecological conditions [2,3]. The importance

of legume crops in agricultural production systems is not limited to their nutritional value; they also attract considerable attention due to their ability to improve soil fertility, reduce energy costs, and support biodiversity. However, environmental stress factors such as drought, heat stress, irregular precipitation patterns, and low soil fertility substantially limit the growth, development, and grain yield of faba bean.

Faba bean yield is shaped by complex interactions among genetic structure, environmental conditions, and cultivation practices. Agronomic traits such as plant height, first pod height, pod length, number of pods per plant, number of seeds per plant, biological yield, and hundred-seed weight are among the principal parameters directly affecting grain yield [4]. Previous studies have demonstrated that traits including flowering time, number of seeds per pod, number of pods per plant, and hundred-seed weight are significantly associated with grain yield. Moreover, genotype × environment interactions become more pronounced in semi-arid regions characterized by high environmental variability, resulting in substantial differences in genotype performance [5]. In a comprehensive study conducted in Türkiye, the evaluation of 372 faba bean genotypes and three commercial cultivars across three different locations over two growing seasons revealed considerable variation in several traits, including grain yield, biological yield, hundred-seed weight, plant height, and flowering time. In the same study, grain yield ranged from 227.14 to 270.78 kg da^{-1}, whereas hundred-seed weight exhibited a broad variation between 62.56 and 231.04 g. These findings indicate that faba bean genetic resources possess significant potential for breeding programs aimed at yield improvement [6].

Yield prediction in agricultural production is of great importance for the selection of suitable genotypes and the optimization of cultivation strategies. Conventional statistical methods, particularly correlation, regression, and path analyses, are widely used to identify yield-related traits [7]. In addition, principal component analysis (PCA) is considered an effective method for interpreting multivariate datasets and determining the main factors affecting yield. Nevertheless, classical statistical approaches may remain insufficient for modeling the nonlinear relationships and complex interactions frequently encountered in agricultural systems. More flexible and powerful modeling approaches are therefore required to fully explain the complex relationships among genotype, environment, and agronomic traits.

In recent years, machine-learning methods have become important tools for prediction and estimation in agricultural research [8,9]. More recently, advanced tabular-learning and foundation-model approaches have attracted attention because they can work with limited-size structured datasets while reducing, but not eliminating, the need for model-specific tuning. In agricultural studies, these approaches may be useful for trait-based yield estimation when directly measured plant variables and biologically motivated derived features are available.

Alongside these developments, explainable-AI methods are increasingly used to interpret complex predictors [10], while deep-learning approaches, including multilayer artificial neural networks (ANNs), have been applied broadly in agricultural analysis and crop-yield modeling [11]. In faba bean, the available yield-estimation and phenotyping studies have more commonly combined UAV-derived RGB, multispectral, thermal, or hyperspectral information with conventional machine-learning or chemometric methods [12–18], indicating that crop-specific deep-learning evidence for faba bean yield remains limited. Evidence from related grain legumes likewise suggests that ANN/XAI frameworks can identify reproductive traits relevant to yield estimation, as reviewed in Section 2. These studies motivate neural approaches, but they do not remove the small-sample constraints of the present dataset. Accordingly, this study evaluates TabPFN as a pretrained model for limited-

size tabular data rather than claiming that it is categorically preferable to conventional deep-learning architectures.

Interpretability remains an important limitation of complex machine-learning models. SHapley Additive exPlanations (SHAP) assigns additive feature attributions to individual predictions and can be summarized to describe global patterns of model reliance [10]. Such attributions explain model behavior but do not establish causal biological effects. This distinction motivates the combined use of SHAP, permutation importance, and LOCO in the present framework.

Although substantial research has been conducted on agronomic characterization, genetic diversity, correlation analyses, and genome-wide association studies (GWAS) in faba bean, studies focusing on trait-based single-plant yield estimation using explainable tabular machine learning and foundation-model-based approaches remain limited. In particular, there is still a need for modeling frameworks that can simultaneously provide accurate estimation, statistical validation, and biological interpretation of yield-related agronomic traits. In this context, the present study aimed to estimate single-plant yield in faba bean genotypes using an explainable TabPFN-based tabular machine learning framework and to identify the agronomic variables driving model predictions through complementary explainability analyses. The novelty of the study lies in integrating a pre-trained tabular foundation model with comparative machine learning models, Holm-corrected statistical testing, and SHAP-, permutation-, and LOCO-based interpretation within a trait-based yield estimation framework.

This study strengthens the use of data-driven and interpretable tabular machine learning approaches in faba bean trait-based yield estimation and contributes to the development of biologically informed decision-support strategies for genotype evaluation and yield-oriented trait prioritization.

Our work contributes to the literature by introducing an interpretable, trait-based machine learning framework for single-plant yield estimation in faba bean:

1. The study applies TabPFN, a pre-trained tabular foundation model, to faba bean single-plant yield estimation, demonstrating the potential of foundation-model-based tabular learning for modeling complex relationships among agronomic, reproductive, and yield-related traits in a limited-size experimental dataset.
2. The modeling framework combines 13 original agronomic measurements with 6 derived yield-related features, allowing both direct plant traits and biologically meaningful composite indicators to be evaluated within the same predictive structure.
3. Unlike many previous faba bean yield studies that mainly focus on remote sensing, image-based phenotyping, or conventional regression-based analysis, this study provides a trait-based yield estimation framework using directly measured agronomic variables.
4. The study compares multiple machine learning regression models under the same validation strategy and supports model evaluation with held-out test-set performance and Holm-corrected Wilcoxon signed-rank tests based on paired absolute errors.
5. The integration of SHAP, permutation importance, and LOCO analyses provides complementary evidence on feature contribution, model sensitivity, and potential redundancy among agronomic variables.
6. By combining predictive modeling, statistical comparison, and biological interpretability, the study offers a data-driven decision-support approach for genotype evaluation, trait prioritization, and yield-oriented selection in faba bean breeding studies.
7. The proposed framework is positioned as a trait-based yield estimation approach using directly measured and derived agronomic traits, rather than an early-stage prediction system.

Direct yield components, including pod weight and biological yield, were included not for pre-harvest forecasting but to quantify predictive attribution among harvest-stage agronomic traits. The framework uses prediction as a computational tool for post-harvest trait ranking; it does not estimate causal effects. The resulting attributions may help formulate candidate multi-trait selection criteria, but any trait weights require independent validation in breeding populations and additional environments.

The article is organized as follows. Section 2 reviews related studies on faba bean yield estimation, machine learning, and explainable AI. Section 3 describes the plant material, field experiment, dataset preparation, preprocessing, exploratory analysis, and feature engineering. Section 4 presents model performance, explainability analyses, and statistical comparisons. Section 5 discusses the findings and their implications for trait prioritization and breeding. Section 6 concludes the study and outlines future research directions.

2. Related Works

Recent advances in machine learning (ML), remote sensing, and high-throughput phenotyping have significantly improved the analysis of agronomic traits and yield-related characteristics in faba bean (*Vicia faba* L.). UAV-based RGB, multispectral, thermal, and hyperspectral sensing technologies have been widely employed to estimate plant height, biomass, canopy development, grain quality, and yield-related parameters. Previous studies demonstrated that machine learning algorithms such as Support Vector Regression (SVR), Random Forest (RF), XGBoost, Partial Least Squares Regression (PLSR), artificial neural networks, and ensemble learning methods can effectively model complex relationships between sensor-derived features and crop performance [12–18]. In many cases, sensor-fusion strategies combining structural, spectral, and thermal information achieved higher prediction accuracy than single-sensor approaches [13,14,16].

Beyond yield estimation, machine learning has also been applied to cultivar identification, seed classification, and grain quality assessment in faba bean. Morphological image analysis, RGB-based seed characterization, near-infrared spectroscopy (NIRS), and chemometric approaches have been successfully used for the prediction of protein content, oil composition, and varietal discrimination [19–21]. In addition, ML and hybrid optimization models have been employed for seasonal yield prediction and processing-related applications, such as drying kinetics and moisture ratio estimation [22,23]. These studies collectively demonstrate the growing importance of data-driven approaches in faba bean phenotyping, quality assessment, and breeding research. However, most of these approaches rely primarily on image-derived, spectral, or remotely sensed information. Comparatively fewer investigations have focused on directly measured agronomic traits collected at the individual plant level, despite their biological relevance for genotype evaluation and yield-oriented selection.

Across crops, machine-learning and ANN models have been used for yield prediction and trait evaluation [24–26]. In grain legumes, explainable or ANN-based studies have examined global yield drivers, pea seed protein, and flowering-related genetic merit in bean [27–29]. ANN models have also related climatic and agronomic traits to rice yield [30]. In chickpea, an ANN-based explainable-AI study identified reproductive traits—particularly seed weight per plant, pod number, and seeds per plant—as influential for grain-yield prediction under semi-arid conditions [31]. Comparable prediction and trait-prioritization applications have been reported in maize, cassava, and linseed [32–35]. Collectively, these studies show the value of data-driven methods for crop improvement, while differences in species, outcomes, and experimental settings mean that their trait rankings cannot be transferred directly to faba bean.

While ML models often provide high predictive performance, their limited interpretability has traditionally restricted their adoption in breeding and agricultural decision-support systems. Consequently, Explainable Artificial Intelligence (XAI) has emerged as a critical research direction in agricultural analytics, aiming to balance predictive performance with model transparency and biological interpretability. Recent studies have increasingly employed SHAP, Local Interpretable Model-Agnostic Explanations (LIME), permutation importance, Sobol analysis, and related interpretability techniques to identify the variables driving model predictions [27,36–38]. In crop-yield prediction studies, SHAP has been successfully used to quantify the influence of climatic, phenological, spectral, and agronomic variables on model outputs [27,36,37]. Similarly, LIME, Sobol analysis, and other sensitivity-based methods have supported the interpretation of crop coefficient estimation and environmental response modeling [38]. Explainability approaches have also supported feature selection, identification of redundant variables, biological interpretation of trait contributions, and improved transparency of ML systems [39–41].

The growing adoption of XAI reflects a broader transition from purely predictive black-box models toward interpretable and biologically meaningful decision-support frameworks. Such approaches are particularly valuable in plant breeding, where understanding the contribution of specific traits can be as important as achieving high predictive accuracy. In this context, XAI methods provide a bridge between computational prediction and biological interpretation by identifying the traits, environmental variables, or spectral indicators that most strongly influence model behavior [27,36–38,42].

Despite the growing adoption of machine learning in agricultural prediction systems, several important gaps remain in the literature. Most faba bean studies have focused on remote sensing, UAV-based phenotyping, hyperspectral analysis, stress detection, or cultivar classification rather than direct trait-based yield estimation using plant-level agronomic measurements. In addition, relatively few studies have combined accurate yield prediction with comprehensive biological interpretation of the underlying yield-determining traits. Although explainable artificial intelligence techniques such as SHAP are increasingly used to improve model transparency, their application to faba bean yield analysis remains limited. Furthermore, foundation models specifically developed for tabular data have recently emerged as promising alternatives to conventional machine learning approaches. Among these, TabPFN leverages a pre-trained probabilistic foundation model designed specifically for tabular datasets and has demonstrated strong performance on small and medium-sized datasets. Nevertheless, its application in agricultural yield prediction remains largely unexplored, and to the best of our knowledge, no previous study has systematically investigated TabPFN for trait-based single-plant yield estimation in faba bean. To address these research gaps, the present study proposes an interpretable machine learning framework for single-plant yield estimation in faba bean using directly measured agronomic traits and biologically meaningful derived features. Unlike previous studies that predominantly rely on image-based, spectral, or sensor-derived information, the proposed framework focuses on plant-level measurements that are directly relevant to genotype evaluation and yield formation. Furthermore, TabPFN is combined with complementary explainability techniques, including SHAP, permutation importance, and Leave-One-Covariate-Out (LOCO) analyses, enabling both accurate prediction and comprehensive biological interpretation of yield-related traits. Consequently, the proposed approach contributes not only to genotype evaluation, trait prioritization, and yield-oriented breeding decisions in faba bean, but also to the growing adoption of interpretable foundation-model-based learning in agricultural research.

3. Materials and Methods

3.1. Plant Material

In this investigation, 372 local faba bean genotypes were selected from a collection of landraces originating from 35 regions in Türkiye (Adana, Amasya, Antakya, Antalya, Artvin, Aydın, Balıkesir, Diyarbakır, Edirne, Elazığ, Erzincan, Eskişehir, Giresun, İzmir, Isparta, İstanbul, Kahramanmaraş, Kars, Kastamonu, Kayseri, Kırklareli, Kırşehir, Kocaeli, Konya, Kütahya, Malatya, Manisa, Mardin, Mersin, Muğla, Samsun, Sinop, Sivas, Şanlıurfa, and Tekirdağ). The material was obtained as seed accessions from the Aegean Agricultural Research Institute National Gene Bank and the International Center for Agricultural Research in the Dry Areas (ICARDA). Filiz-99, Kıtık-2003, and Salkım were included as standard check cultivars. Information on the broader source collection is provided in Supplementary Table S1.

3.2. Field Experimentation

Although the broader germplasm collection was evaluated at Adana, İzmir, and Sivas during the 2016–2017 and 2017–2018 growing seasons, the dataset analyzed in this study was derived exclusively from the field experiment conducted at Menemen, İzmir, during the 2017–2018 growing season. The experiment was established using an augmented block design comprising nine blocks. Each of the 372 local genotypes was evaluated in one unreplicated plot, whereas the three check cultivars (Filiz-99, Kıtık-2003, and Salkım) were planned once in each block. Thus, the planned design contained nine replicate plots per check cultivar and 27 check plots in total. The final modeling dataset contained 26 observed check plots; therefore, one of the 27 planned check plots was not represented in the analyzed records. Each plot consisted of two rows, each 2 m long, with 0.45 m between rows and 0.10 m between plants, corresponding to a nominal plot area of 1.80 m^2. Twenty seeds were sown in each row, resulting in 40 seeds per plot at establishment. These plants constituted the within-plot population and were not treated as independent experimental replicates; the plot was considered the experimental unit. Fertilizer was applied at rates of 40 kg ha^{-1} N and 80 kg ha^{-1} P_2O_5, equivalent to 4 and 8 kg da^{-1}, respectively. Seeds were sown on 26 November 2017, and the plots were harvested during June 2018. During flowering, the plots were covered with gauze to limit insect-mediated cross-pollination, and appropriate insect-control measures were applied.

The field experiment was conducted at the Aegean Agricultural Research Institute in Menemen, İzmir, Türkiye (38°33′ N, 27°03′ E; approximately 9 m a.s.l.). Pre-sowing soil analysis indicated a clay-loam texture (37.3% sand, 28.0% silt, and 34.7% clay), pH 7.15, electrical conductivity 0.17 dS m^{-1}, organic matter 3.5%, calcium carbonate 4.6%, available P_2O_5 24.8 kg ha^{-1}, and exchangeable K_2O 306 kg ha^{-1}. During the November–May growing period, total precipitation was 432.5 mm, including 221.0 mm in January. The mean seasonal temperature was 11.8 °C; the recorded temperature extremes were −4.2 °C in January and 33.5 °C in May, and mean relative humidity was 64.0%.

3.3. Dataset Collection and Preprocessing

The dataset used in this study was compiled from plot-level agronomic measurements collected during the 2017–2018 İzmir field experiment according to the IBPGR/ICARDA descriptor list for faba bean [43]. Phenological observations were recorded from emergence to physiological maturity, whereas morphological and yield-related traits were measured during reproductive development or after harvest. The final modeling dataset comprised 398 distinct plot-level observations. The 372 local genotypes were represented by one plot each, and the remaining observations represented repeated check plots distributed across the nine augmented blocks. Each observation corresponds to an experimental plot

rather than to an individual plant. Plot-level values, including the target variable single-plant yield (g plant^{-1}), were calculated as arithmetic means of the plants sampled within that plot.

The dataset consisted of 19 predictor variables, including 13 directly measured agronomic traits and 6 derived variables calculated from the original measurements. Days to flowering was recorded as the number of days from sowing until 50% of the plants within the corresponding experimental plot had flowered. Days to maturity was determined as the number of days from sowing until the crop reached physiological maturity. Plant height was measured from the soil surface to the tip of the main stem, while first pod height was measured from the soil surface to the first pod attachment. Branch number, pod number per plant, seed number per pod, and seed number per plant were determined by direct counting using ten randomly selected plants from each experimental plot. Pod length was measured from ten randomly selected mature pods using a ruler. Pod weight, biological yield, and single-plant yield were determined using a precision balance after harvest. Hundred-seed weight was calculated as the mean weight of four independent samples of 100 seeds. Plot yield was calculated from the harvested seed weight of each experimental plot and expressed as kg da^{-1}.

The remaining six explanatory variables were mathematically derived from the original agronomic measurements to improve biological interpretation and prediction performance. These included pod filling rate, seed-to-pod ratio, seed unit weight, pod length–yield index, plant yield ratio, and vegetation period. Their mathematical formulations are presented in Section 3.5.

Prior to model development, the dataset was examined for missing values, inconsistent records, and duplicated observations; no missing values requiring imputation were detected. To prevent data leakage, the dataset was split into training and held-out test subsets before any preprocessing or scaling steps were applied. The splitting process was performed using the train_test_split function with an 80% training and 20% test ratio, and random_state = 42 was set to ensure reproducibility. As a result of this process, 318 observations were included in the training set, while 80 observations were included in the held-out test set. The test set was not used in any model development phase, including hyperparameter optimization, cross-validation, or model selection; it was reserved solely for the held-out evaluation of the final model's performance.

During the preprocessing stage, continuous variables were transformed using the QuantileTransformer implemented in the scikit-learn 1.7.2 library with the parameters output_distribution = 'normal', n_quantiles = 200, and random_state = 42, thereby reducing distributional skewness and improving model stability. The same transformation fitted on the training data was subsequently applied to the held-out test set to prevent data leakage.

Mathematically, for the j-th input feature and the i-th observation, the quantile transformation with normal output distribution [44,45] can be expressed as:

$$\widetilde{x}_{ij} = \Phi^{-1}\left(\hat{F}_j\left(x_{ij}\right)\right), \tag{1}$$

where $\widetilde{x}_{ij}$ denotes the original value of the j-th feature for the i-th observation, $\hat{F}_j$ is the empirical cumulative distribution function estimated from the training data for the j-th feature, and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution.

The QuantileTransformer reduces the impact of outliers by approximating each variable's marginal distribution with a standard normal distribution, thereby making the variable distributions more suitable for model training. This approach was chosen specifically to improve model stability in tabular datasets where agronomic variables with different scales and distribution structures are used together.

Scaling and transformation parameters were learned exclusively on the training dataset, and the transformation learned from the training set was directly applied to the test dataset. This prevented indirect information transfer from the test set to the training process and avoided potential data leakage arising from the preprocessing stage. This approach enabled a more realistic and reliable assessment of model performance and ensured that the held-out test set served as an unbiased evaluation criterion for measuring the model's generalization ability.

3.4. Exploratory Data Analysis

Before the feature engineering and model development stages, exploratory data analysis was conducted to assess the overall structure, level of variability, and basic distribution characteristics of the original agronomic dataset. At this stage, only directly measured agronomic traits were considered; derived traits were not included in this assessment because they were generated during the subsequent feature engineering process. In the exploratory Principal component analysis (PCA) and Pearson correlation analyses, only the original measured agronomic traits were used in order to describe the raw structure of the dataset before feature engineering. In contrast, the machine learning regression models were trained using the full predictor set, consisting of 13 original agronomic traits and 6 derived features. Thus, the PCA/correlation analyses and the regression models were based on related but not identical feature sets. Descriptive statistics for the original agronomic traits are presented in Table 1.

Table 1. Descriptive statistics of the original agronomic traits used in the exploratory analysis.

No	Feature	Mean	SD	Min	Max	CV (%)
1	Flowering Duration (day)	47.84	2.07	40.00	52.00	4.33
2	Days to Maturity (day)	106.18	2.87	99.00	112.00	2.70
3	First Pod Height (cm)	31.79	6.92	14.80	57.60	21.76
4	Branch Number (plant^{-1})	2.95	0.70	1.54	7.80	23.84
5	Plant Height (cm)	75.15	11.79	49.20	107.80	15.69
6	Biological Yield (g plant^{-1})	34.08	12.07	11.17	105.17	35.41
7	Pod Number (plant^{-1})	7.55	2.55	1.60	16.40	33.78
8	Pod Weight (g)	17.70	6.87	2.83	44.52	38.77
9	Pod Length (cm)	7.46	1.04	4.80	11.40	13.89
10	Seed Number per Pod	2.80	0.92	1.31	14.80	32.97
11	Seed Number per Plant	14.71	5.61	2.40	40.40	38.10
12	Plot Yield (kg da^{-1})	191.60	93.85	28.00	515.45	48.98
13	100-Seed Weight (g)	119.34	21.32	44.58	167.63	17.87

The exploratory data analyses were conducted using a combination of statistical and programming software environments. PCA was performed using JMP® version 7.0 (SAS Institute Inc., Cary, NC, USA). Prior to PCA, the 13 original agronomic traits were standardized to zero mean and unit variance to eliminate differences arising from measurement scales, so that the analysis was effectively based on the correlation matrix of the traits. The Pearson product–moment correlation coefficients were computed in Python 3.13.0 (Python Software Foundation, Wilmington, DE, USA) under Windows 11 (64-bit) with pandas 2.3.2 using the DataFrame.corr (method = 'pearson') function, and the corresponding coefficients and two-tailed significance levels were independently verified with SciPy 1.16.2 (scipy.stats.pearsonr). Numerical computations were performed with NumPy 2.3.3, and the correlation heat map was generated with Matplotlib 3.10.6 and seaborn 0.13.2.

The coefficient of variation (CV) indicated significant differences in variability among the traits examined. Phenological traits generally exhibited low variation. The flowering period ranged from 40.00 to 52.00 days, and the CV value was calculated as 4.33%. Similarly, the maturation period fell within the range of 99.00–112.00 days and was the trait showing the lowest variability, with a CV value of 2.70%. These results indicate that the genotypes evaluated exhibit relatively limited variation in terms of their phenological development periods.

In contrast, a higher degree of variability was observed in traits related to yield and yield components. The highest CV value was observed for plot yield; this trait ranged from 28.00 to 515.45 kg da^{-1} and had a CV of 48.98%. Pod weight (38.77%), number of seeds per plant (38.10%), biological yield (35.41%), number of pods per plant (33.78%), and number of seeds per pod (32.97%) also exhibited high variation. These findings indicate significant differences among genotypes in terms of reproductive capacity, pod development, seed formation, and biomass accumulation.

Morphological traits, however, showed moderate variability. The CV values for first pod height, number of branches, and plant height were determined as 21.76%, 23.84%, and 15.69%, respectively. Pod length and 100-seed weight, on the other hand, exhibited more limited but still notable variation with CV values of 13.89% and 17.87%, respectively. Overall, descriptive statistics indicate that phenological traits are more stable, whereas pod, seed, and biomass traits, which are directly related to yield formation, are more variable. This variation pattern provides a biologically meaningful data foundation for subsequent modeling stages aimed at estimating single-plant yield.

Since single-plant yield is the primary variable targeted for estimation in this study, the distribution structure was also examined before modeling. As shown in Figure 1, the majority of single-plant yield values are concentrated in the medium yield range, whereas a limited number of observations with high yield values were identified. This distribution indicates that there is sufficient variation in single-plant yield within the dataset, and that it provides a biologically meaningful basis for the subsequent prediction-based modeling phase.

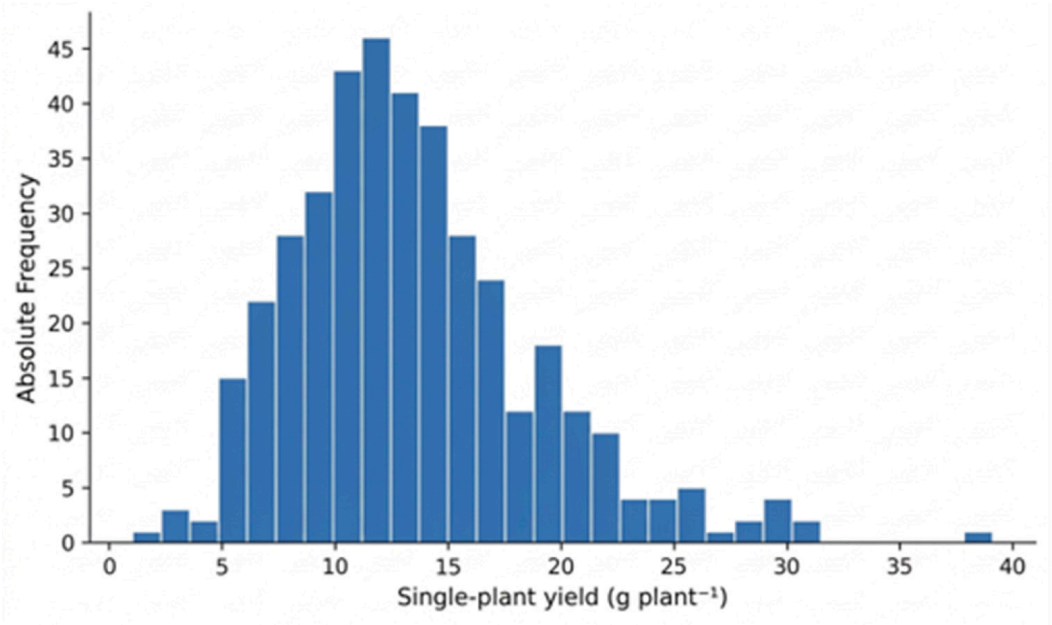

Figure 1. Distribution of single-plant yield values in the original dataset before feature engineering and model development.

Principal Component Analysis

Principal component analysis (PCA) was performed to summarize the multivariate structure of the original agronomic dataset and to identify the principal sources of variation among the evaluated traits. Only the 13 directly measured agronomic variables were included in the PCA; the six mathematically derived variables were excluded because they were generated during the subsequent feature-engineering stage. Since the agronomic variables were expressed in different units and measurement scales, the data were standardized before analysis. PCA was conducted using JMP® version 7.0 (SAS Institute Inc., Cary, NC, USA). Principal components were derived from the correlation matrix of the standardized variables. Component scores were used to evaluate the distribution of the observations, while trait loadings were used to examine the contribution of individual variables to the principal components. The PCA results, including the distribution of observations and relationships among agronomic traits, are presented and biologically interpreted in the Results section.

3.5. Feature Engineering

In addition to the 13 directly measured agronomic traits, six study-specific features were constructed using domain-informed feature engineering. Biological knowledge can guide transformations of original predictors into ratios, differences, or composite variables that represent selected relationships more explicitly [46]. In faba bean, yield has been associated with pod and seed traits, biomass, seed weight, and phenological development [1,5–7]. These studies provide agronomic context for evaluating composite predictors; they do not establish the six formulas used here as standardized indices.

Accordingly, six derived variables were defined to represent specific relationships among the original agronomic measurements: pod filling rate (pod weight/pod number), seed-to-pod ratio (seed number per plant/pod number), seed unit weight (pod weight/seed number per plant), pod length–yield index (pod weight/pod length), plant yield ratio (biological yield/pod number), and vegetation period (days to maturity − days to flowering). These variables were designed to represent, respectively, average pod filling, seed production per pod, weight allocation per seed, pod filling relative to pod size, biomass production per pod, and the duration between flowering and maturity.

The six formulas and their names were defined specifically for the present modeling framework and should not be interpreted as previously validated or standardized agronomic indices. Their incremental predictive contribution was evaluated empirically in the ablation study (Section 4.7).

The six derived features were denoted as follows: pod filling rate (PFR), seed-to-pod ratio (SPR), seed unit weight (SUW), pod length–yield index ($PLYI$), plant yield ratio (PYR), and vegetation period (VP). These variables were calculated at the observation level to represent biologically meaningful relationships among pod development, seed formation, biomass accumulation, and phenological duration [46].

$$PFR_i = \frac{PodWeight_i}{PodNumber_i + \varepsilon}, \quad (2)$$

$$SPR_i = \frac{SeedsPerPlant_i}{PodNumber_i + \varepsilon}, \quad (3)$$

$$\mathrm{SUW}_i = \frac{\mathrm{PodWeight}_i}{\mathrm{SeedsPerPlant}_i + \varepsilon}, \quad (4)$$

$$\mathrm{PLYI}_i = \frac{\mathrm{PodWeight}_i}{\mathrm{PodLength}_i + \varepsilon}, \quad (5)$$

$$\mathrm{PYR}_i = \frac{\mathrm{BiologicalYield}_i}{\mathrm{PodNumber}_i + \varepsilon}, \tag{6}$$

$$\mathrm{VP}_i = \mathrm{DaysToMaturity}_i - \mathrm{DaysToFlowering}_i, \tag{7}$$

where i denotes the i-th observation and $\varepsilon = 0.01$ was used in the ratio-based features to avoid denominator instability. The derived variables were not used in the exploratory PCA and Pearson correlation analyses, which were based only on the original measured agronomic traits, but they were included in the machine learning regression models as part of the full 19-variable predictor set.

3.6. Computational Environment and Model Efficiency

All computational experiments, including model development, hyperparameter optimization, repeated-split evaluation, and explainability analyses, were conducted on a Lenovo LOQ 15IRX9 laptop computer (Lenovo Group Ltd., Beijing, China) equipped with a 20-thread Intel Core i7-13650HX processor (Intel Corporation, Santa Clara, CA, USA) and an NVIDIA GeForce RTX 4050 Laptop GPU (6 GB; NVIDIA Corporation, Santa Clara, CA, USA), running Windows 11 (64-bit; Microsoft Corporation, Redmond, WA, USA). The analysis environment comprised Python 3.13.0, PyTorch 2.7.1 with CUDA 11.8, tabpfn 7.1.1, scikit-learn 1.7.2, XGBoost 3.1.2, and Optuna 4.8.0. TabPFN was executed on the GPU, whereas the five baseline models were executed on the CPU. On the reference split, fitting TabPFN on 318 training observations required 0.226 s and predicting the 80 held-out observations required 0.288 s. The five baseline searches together required approximately 1814 s because each combined 100 Optuna trials with five-fold cross-validation. Runtime values are hardware dependent and are therefore reported as computational-profile measurements rather than hardware-neutral model properties (Supplementary Table S3).

3.7. Machine Learning Models and Evaluation Metrics

Six regression algorithms were selected to span complementary modeling paradigms for limited-size tabular data: Random Forest and Extra Trees represent tree-based bagging; XGBoost and HistGradientBoosting represent gradient boosting; ν-SVR provides a nonlinear kernel baseline; and TabPFN represents a pretrained transformer-based tabular foundation model. Comparing these model families under the same split and validation protocol allows their different inductive biases to be assessed on the present dataset. The selection does not imply that any family is universally optimal for agricultural regression. TabPFN was the focal model because the study aimed to test whether its learned prior could achieve competitive accuracy with less dataset-specific optimization than the tuned conventional baselines.

To ensure a fair comparison, the five conventional baseline models (Random Forest, Extra Trees, Support Vector Regression, HistGradientBoosting, and XGBoost) were optimized with Optuna 4.8.0 using the Tree-structured Parzen Estimator sampler with seed = 42 and 100 trials per model. Each trial was evaluated by five-fold cross-validation on the training set, with mean R^2 across the folds as the optimization objective. In every trial, the model and the QuantileTransformer were combined in a scikit-learn 1.7.2 Pipeline so that the transformer was fitted only on the training portion of each fold. Support Vector Regression was implemented as ν-SVR (NuSVR) with a radial basis function kernel; the tuned parameters were therefore ν, C, and γ. TabPFN was not included in the Bayesian optimization. Only n_estimators was examined over {4, 8, 16, 32}, and n_estimators = 4 was selected. The complete search spaces and selected configurations are reported in Supplementary Table S2.

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of

Determination (R^2), which are standard regression evaluation metrics [44]. These four metrics are calculated as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{9}$$

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}. \tag{11}$$

Here, y_i is the observed single-plant yield, $\hat{y}_i$ is the predicted value, $\overline{y}$ is the mean of the observed values, and N is the number of observations. Since all observed single-plant yield values in the dataset were positive, the MAPE was well-defined for the present analysis.

TabPFN was implemented using the official TabPFN Python package (version 7.1.1), which provides the second-generation tabular foundation model architecture. Specifically, the tabpfn.regressor.TabPFNRegressor class was executed with a PyTorch 2.7.1 (CUDA 11.8) backend. The pre-trained model checkpoint distributed with the package was used directly without dataset-specific fine-tuning. Only the number of ensemble members (n_estimators) was evaluated, while all remaining parameters were retained at package defaults (softmax_temperature = 0.9, average_before_softmax = False, fit_mode = 'fit_preprocessors', inference_precision = 'auto'). The random seed was fixed at random_state = 42 for exact reproducibility. All input features were transformed using the QuantileTransformer described in Section 3.3 prior to model input.

3.8. Explainability and Statistical Analyses

A SHAP-based explainability analysis was conducted to interpret the prediction mechanism of the TabPFN model. Global SHAP importance was calculated as the mean absolute SHAP value for each feature, following the SHAP framework of [47], as follows:

$$I_j = \frac{1}{N}\sum_{i=1}^{N}|\varphi_{ij}|, \tag{12}$$

$$\mathrm{RC}_j = 100 \times \frac{I_j}{\sum_{k=1}^{s} I_k}, \tag{13}$$

$$\mathrm{CC}_j = \sum_{\ell=1}^{j}\mathrm{RC}_{(\ell)}, \tag{14}$$

Here, I_j denotes the mean absolute SHAP importance of the j-th feature, φ_{ij} is the SHAP value of the j-th feature for the i-th observation, N is the number of observations, and s is the total number of input features. The relative contribution RC_j expresses the percentage contribution of the j-th feature to the total SHAP importance. The cumulative contribution CC_j is computed after ranking the features in descending order of I_j, where $\mathrm{RC}_{(\ell)}$ denotes the relative contribution of the ℓ-th ranked feature.

Permutation importance analysis was performed on the trained TabPFN model using the held-out test set. For each input variable, the values of that variable were randomly shuffled while all other variables were kept unchanged, and the resulting decrease in model performance was measured using the coefficient of determination (R^2). This procedure

was repeated 50 times for each feature, and the mean R^2 decrease and standard deviation were calculated.

The Leave-One-Covariate-Out (LOCO) analysis was conducted to evaluate the unique predictive contribution of each input variable when the model was allowed to relearn without that variable. For this purpose, each input variable was sequentially removed from the predictor set, the TabPFN model was retrained using the remaining variables, and the resulting test-set R^2 value was compared with the baseline model performance.

Finally, pairwise differences in predictive performance were assessed using the Wilcoxon signed-rank test. For each pair of models, the test was applied to the paired absolute prediction errors obtained for the same observations in the held-out test set. To control the family-wise Type I error rate arising from the 15 pairwise comparisons among the six models, the resulting p-values were adjusted using the Holm step-down procedure. Statistical significance was evaluated at $\alpha = 0.05$. For models A and B, the paired absolute-error difference was defined as:

$$d_i = \left|y_i - \hat{y}_i^{(A)}\right| - \left|y_i - \hat{y}_i^{(B)}\right|. \tag{15}$$

For m pairwise comparisons, the Holm step-down criterion is given by [48]:

$$p_{(k)} \leq \frac{\alpha}{m - k + 1}, \qquad k = 1, \ldots, m. \tag{16}$$

where d_i represents the difference between the paired absolute prediction errors of models A and B for the i-th test observation, y_i denotes the observed single-plant yield value, and $\hat{y}_i^{(A)}$ and $\hat{y}_i^{(B)}$ denote the predicted values produced by models A and B, respectively. In the Holm step-down criterion, m is the total number of pairwise comparisons, $p_{(k)}$ denotes the k-th smallest unadjusted p-value, and α is the predefined significance level.

To complement the pairwise analyses with an omnibus comparison of all six models, the Friedman test was applied to the same matrix of absolute prediction errors for the 80 held-out test observations. Each test observation was treated as a block, and the six models were ranked within each observation from the lowest to the highest absolute prediction error; tied values, if present, were assigned average ranks. Following rejection of the omnibus null hypothesis of equal model performance, pairwise comparisons of the mean ranks were conducted using the Nemenyi post-hoc procedure at $\alpha = 0.05$, following Demšar [49]. The Nemenyi critical difference was calculated as $q_\alpha \sqrt{\frac{K(K+1)}{6N}}$, where q_α is the Studentized-range critical value for K models, $K = 6$ is the number of models, and $N = 80$ is the number of test observations.

The Friedman analysis was supplemented with the Iman–Davenport F approximation and Kendall's W as an omnibus effect-size measure. The reference-split Wilcoxon–Holm and Friedman–Nemenyi results reported in Section 4.6 were computed from the same 80-observation $\times$ 6-model absolute-error matrix, without resampling, aggregation, or exclusion of held-out observations.

Sensitivity to the random train/test partition was assessed using 10 repeated 80/20 splits with seeds 0–9. The baseline hyperparameters selected on the reference split (seed = 42; 100 Optuna trials) were held fixed and each model was refitted on every split; TabPFN was likewise refitted with n_estimators = 4 and without a new per-split hyperparameter search. Test metrics were summarized as mean $\pm$ SD, together with the number of splits in which each model ranked first. Paired Wilcoxon signed-rank tests compared TabPFN with each baseline across the 10 repeated splits, and Holm correction was applied across the five comparisons. Because these splits reuse observations and are not statistically

independent external samples, the repeated-split *p*-values were interpreted as robustness diagnostics rather than as substitutes for external validation.

4. Results

4.1. Principal Component Analysis and Relationships Among Agronomic Traits

Principal component analysis revealed considerable multivariate variation among the evaluated faba bean observations. The distribution of the observations across the principal-component space reflected the combined effects of phenological development, plant architecture, biomass production, pod formation, and seed-related characteristics. The loading patterns indicated that reproductive and biomass-related traits contributed more strongly to the major axes of variation, whereas several phenological traits showed comparatively weaker associations with the main yield-related gradients.

Examination of the scatter plot between maturity duration and flowering duration revealed considerable variation among the evaluated genotypes (Figure 2). However, the relatively homogeneous distribution of data points across the graph indicates the absence of a strong linear relationship between these two phenological traits. Genotypes exhibiting longer flowering periods were distributed across a broad range of maturity durations, suggesting that phenological development follows different strategies among genotypes. This finding indicates that maturity duration is influenced not only by flowering duration but also by genetic background and environmental conditions.

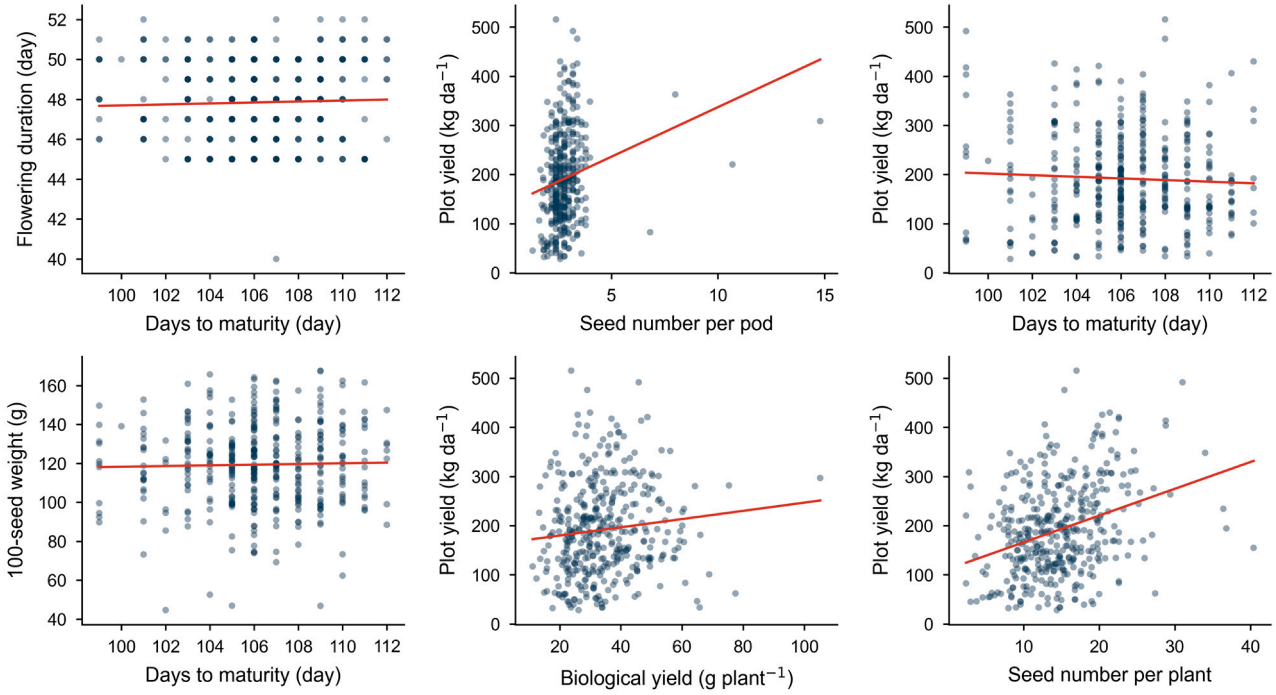

Figure 2. Scatter plots showing the relationships among selected agronomic traits and plot yield.

The relationship between seed per pod and grain yield exhibited a generally positive trend. An increase in the number of seeds per pod was associated with higher grain yield (Figure 2). Nevertheless, the presence of substantial yield variation among genotypes with similar seed numbers per pod suggests that yield cannot be explained by this trait alone. These findings confirm the complex nature of yield as a quantitative trait controlled by multiple interacting components. The concentration of high-yielding genotypes at higher seed numbers per pod indicates that this trait may serve as a valuable selection criterion in breeding programs.

The scatter plot between seed numbers per plant and grain yield demonstrated one of the strongest positive associations among the evaluated traits (Figure 2). Grain yield increased substantially with increasing seed number per plant. Many high-yielding genotypes were clustered within the upper range of seed number values, indicating that this trait is one of the major determinants of yield formation. However, differences in yield among genotypes with similar seed numbers suggest that other characteristics, such as seed weight and biological yield, also contribute significantly to final grain yield.

A clear positive relationship was observed between biological yield and grain yield (Figure 2). Genotypes with higher biological yield generally exhibited superior grain yield performance. However, several genotypes produced relatively high biological yield while maintaining only moderate grain yield levels, indicating differences in assimilate partitioning efficiency from vegetative biomass to reproductive organs. This observation highlights the importance of harvest index in determining final yield performance. Overall, the results demonstrate that biological yield is a major contributor to grain yield formation.

The relationship between maturity duration and grain yield appeared relatively weak, with no obvious linear trend observed across the evaluated genotypes (Figure 2). Both high- and low-yielding genotypes were found within similar maturity periods, suggesting that maturity duration alone is not a direct determinant of yield performance. Nevertheless, some late-maturing genotypes exhibited comparatively higher yield values, implying that extended growth periods may enhance biomass accumulation and yield potential under favorable environmental conditions.

The association between maturity duration and 100-seed weight was generally weak. The wide dispersion of data points indicates that seed size is predominantly under genetic control and can vary independently of maturity duration. Although certain late-maturing genotypes tended to exhibit higher 100-seed weights, this trend was not consistently observed across all genotypes. Therefore, maturity appears to have only a limited influence on seed weight determination.

When all scatter plots were evaluated collectively, the strongest relationships with grain yield were observed for seed number per plant, seed number per pod, and biological yield. In contrast, phenological traits such as maturity duration, flowering duration, and 100-seed weight exhibited relatively weaker and more indirect associations with yield performance. These findings indicate that grain yield is primarily governed by reproductive components and biomass production capacity rather than by phenological characteristics alone.

From a breeding perspective, traits such as seed number per plant and biological yield should be considered among the primary selection criteria for identifying high-yielding genotypes. The observed relationships are also biologically consistent with the findings reported in recent machine learning and SHAP-based analyses, which identified reproductive traits as the most influential determinants of grain yield.

Before modeling, a Pearson correlation analysis was conducted to assess the linear relationships among the original agronomic traits. As shown in Figure 3, the most pronounced positive relationships were observed between the number of seeds per plant and the number of pods per plant, between the number of seeds per plant and pod weight, and between biological yield and pod weight. These results indicate that the variation in yield within the raw dataset is largely associated with pod formation, seed number, and biomass accumulation. The weaker relationships between phenological traits and yield components suggest that, in the current dataset, phenological differences may play a more limited explanatory role compared to yield components.

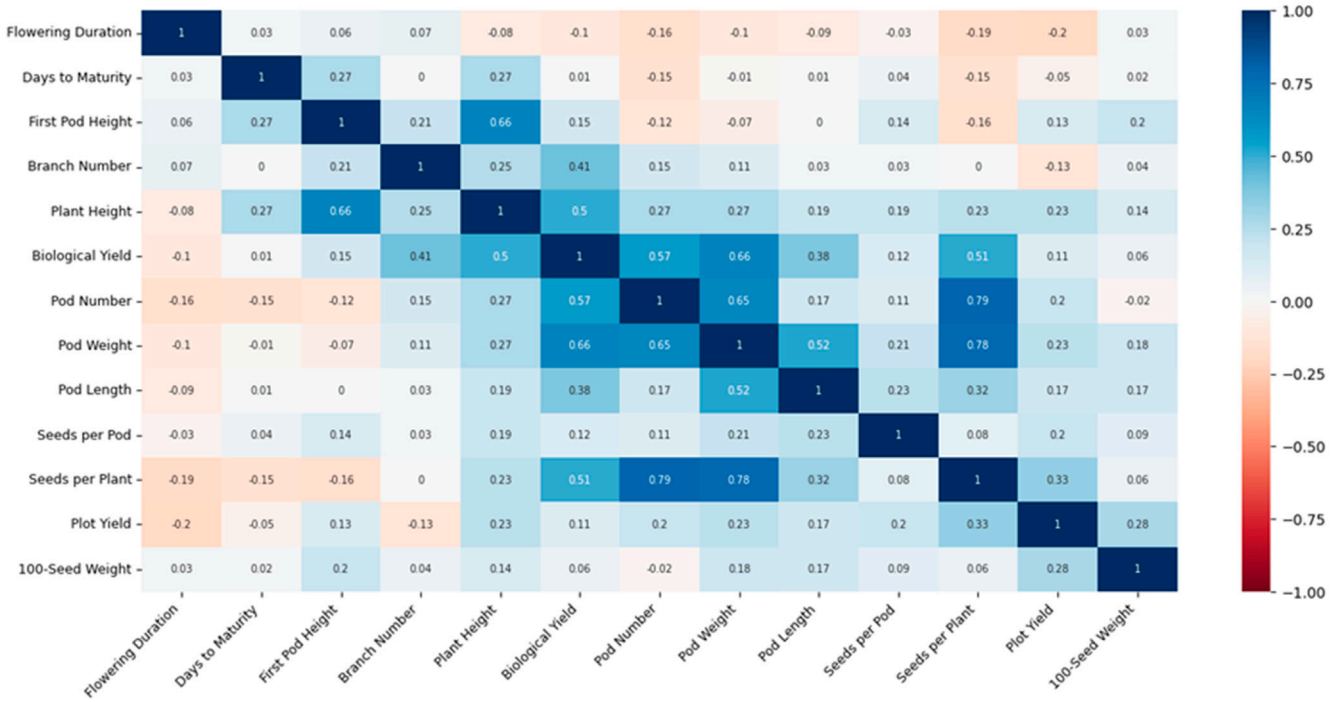

Figure 3. Pearson correlation matrix of original agronomic traits before feature engineering and model development.

To quantify the mathematical and biological dependence among the evaluated predictors, pairwise Pearson correlations and variance inflation factors (VIF) were calculated on the training set (n = 318). As detailed in Table 2, substantial collinearity was observed among several original and derived variables. Most notably, pod weight exhibited near-collinearity with the pod length-yield index (r = 0.929; VIF = 86.3 and 65.4), while days to maturity, flowering duration, and vegetation period formed an exactly collinear triplet (VIF $\rightarrow \infty$) due to their algebraic identity.

Table 2. Feature-dependence diagnostics for the predictor set (training set, n = 318): trait pairs with $|r| \geq 0.70$ and the predictors with the highest variance inflation factors.

Trait Pair/Predictor	Pearson r	VIF	Comment
Pod Weight–Pod Length-Yield Index	0.929	86.3/65.4	Derived feature is pod weight/pod length
Pod Number–Seeds per Plant	0.785	15.9/34.1	Biologically coupled yield components
Pod Weight–Seeds per Plant	0.779	86.3/34.1	Biologically coupled yield components
Seeds per Plant–Pod Length-Yield Index	0.770	34.1/65.4	Shared pod-weight component
Pod Filling Rate–Seed-to-Pod Ratio	0.706	27.5/18.2	Both derived per-pod ratios
Days to Maturity–Vegetation Period	0.806	∞	Exact identity: vegetation = maturity − flowering
Flowering Duration	–	∞	Third member of the exactly collinear triplet

Note: VIF, variance inflation factor, computed as $1/(1 - R^2)$ from the regression of each predictor on all remaining predictors (training set, n = 318). For the pairs, the two VIF values are given in the order of the traits listed in the first column. VIF $\rightarrow \infty$ indicates exact linear dependence.

4.2. Model Performance Evaluation

The predictive performance of the six machine-learning models on the reference split is summarized in Table 3.

Table 3 jointly reports five-fold cross-validation performance on the training set and performance on the held-out test set. TabPFN produced the best numerical result for all four held-out metrics, with the highest R^2 (0.8746) and the lowest RMSE (1.9132 g plant^{-1}), MAE (1.0819 g plant^{-1}), and MAPE (8.16%). The tuned baselines were closely

grouped: the test-R^2 difference between the second-ranked SVR (0.8585) and the sixth-ranked HistGradientBoosting model (0.8458) was only 0.0127.

Table 3. Performance comparison of ML models for single-plant yield estimation in faba bean on the reference split.

Model	CV R^2 (±SD)	CV RMSE (g plant^{-1})	CV MAE (g plant^{-1})	CV MAPE (%)	Test R^2	Test RMSE (g plant^{-1})	Test MAE (g plant^{-1})	Test MAPE (%)
TabPFN	0.8660 (±0.0803)	1.9350 (±0.6755)	1.0528 (±0.1565)	8.35 (±0.86)	0.8746	1.9132	1.0819	8.16
SVR	0.8360 (±0.0848)	2.1426 (±0.6879)	1.1651 (±0.2103)	9.24 (±1.72)	0.8585	2.0326	1.2618	9.53
XGBoost	0.8420 (±0.0904)	2.1076 (±0.6964)	1.2563 (±0.1942)	10.79 (±2.16)	0.8562	2.0485	1.2297	9.09
RandomForest	0.8460 (±0.0913)	2.0756 (±0.7064)	1.2116 (±0.2217)	10.39 (±2.20)	0.8533	2.0692	1.2607	9.59
ExtraTrees	0.8543 (±0.0879)	2.0184 (±0.6970)	1.1499 (±0.2068)	9.90 (±2.04)	0.8497	2.0944	1.2358	9.30
HGB	0.8391 (±0.0880)	2.1320 (±0.6840)	1.3064 (±0.2031)	11.77 (±2.65)	0.8458	2.1213	1.3566	10.22

Note: reference split, random_state = 42. CV denotes five-fold cross-validation on the training set (n = 318); Test denotes the held-out test set (n = 80). R^2 is dimensionless; RMSE and MAE are expressed in g plant^{-1}; MAPE is a percentage. CV entries are the mean across five folds, with the fold standard deviation in parentheses. Models are ordered by descending test R^2. The five baseline models were optimized with 100 Optuna trials (TPE sampler, seed = 42); TabPFN did not undergo Optuna optimization, and n_estimators = 4 was selected from {4, 8, 16, 32}.

The fold-to-fold standard deviations of cross-validation R^2 ranged from 0.0803 to 0.0913, which were substantially larger than the observed between-model differences in mean R^2; the precise model ordering should therefore be interpreted cautiously and together with the statistical comparisons and repeated-split analysis in Sections 4.6 and 4.8. The cross-validation means and held-out test values were generally similar, indicating no obvious large generalization gap on the reference split, although this agreement alone does not rule out overfitting. The partial disagreement between RMSE- and MAPE-based rankings—for example, Extra Trees had a lower MAPE but a higher RMSE than Random Forest—is consistent with the different weighting properties of these metrics: RMSE gives greater weight to large absolute residuals, whereas MAPE evaluates error relative to the observed yield.

4.3. SHAP-Based Explainability Analysis

A SHAP-based explainability analysis was conducted to interpret the prediction mechanism of the TabPFN model and identify the variables with the greatest influence on single-plant yield. SHAP importance values were calculated using mean absolute SHAP values, which represent the average absolute contribution of each variable to the model's predictions.

A higher SHAP importance value indicates that the corresponding variable has a stronger average contribution to the model output. To quantitatively detail this hierarchy, the exact global importance rankings, along with their relative contribution ratios and cumulative contribution percentages, which are essential for identifying the cumulative impact of top traits, are presented in Table 4 below.

According to the SHAP results, the variable that most influenced the model predictions was pod weight. With a SHAP importance value of 1.1677, pod weight alone accounted for 21.52% of the total contribution. This finding indicates that pod weight is one of the key determinants in predicting single-plant yield. Ranking second, the number of seeds per

plant made a strong contribution to the model output with a SHAP importance value of 1.0479 and a relative contribution rate of 19.32%. Pod weight and the number of seeds per plant together accounted for 40.84% of the total model importance, highlighting the central role of reproductive capacity and seed formation in yield prediction.

Table 4. Global feature importance ranking based on mean absolute SHAP values, relative contribution ratios, and cumulative contribution values.

Rank	Feature (Input Variable)	SHAP Importance	Relative Contribution (%)	Cumulative Contribution (%)
1	Pod Weight (g)	1.1677	21.52%	21.52%
2	Seeds per Plant	1.0479	19.32%	40.84%
3	Pod Length-Yield Index	0.8933	16.47%	57.31%
4	Pod Filling Rate	0.3541	6.53%	63.83%
5	Biological Yield (g)	0.3466	6.39%	70.22%
6	Pod Number per Plant	0.3161	5.83%	76.05%
7	Seed Unit Weight	0.3068	5.66%	81.70%
8	Pod Length (cm)	0.2459	4.53%	86.23%
9	Seed-to-Pod Ratio	0.2279	4.20%	90.43%
10	Seeds per Pod	0.1403	2.59%	93.02%
11	Plant Yield Ratio	0.1207	2.23%	95.25%
12	Days to Maturity	0.0612	1.13%	96.37%
13	Days to Flowering	0.0578	1.07%	97.44%
14	Plot Yield (kg da^{-1})	0.0440	0.81%	98.25%
15	Branch Number	0.0295	0.54%	98.80%
16	Plant Height (cm)	0.0260	0.48%	99.27%
17	First Pod Height (cm)	0.0190	0.35%	99.62%
18	Vegetation Period (days)	0.0138	0.25%	99.88%
19	100-Seed Weight (g)	0.0066	0.12%	100.00%

The third most important variable was the pod length-yield index. This variable accounted for 16.47% of the total SHAP contribution, with a SHAP importance value of 0.8933. Together, the top three variables reached a cumulative contribution of 57.31%, indicating that more than half of the total SHAP-based importance was attributed to pod weight, number of seeds per plant, and pod length-yield index. The fact that all three variables are related to pod development, seed formation, and pod filling structure supports the biological interpretability of the TabPFN model.

The first three variables were followed by pod filling rate (6.53%) and biological yield (6.39%). Together, the first five variables explained 70.22% of the total SHAP contribution. This indicates that the TabPFN model primarily utilizes traits related to pod development, seed formation, and biomass accumulation when estimating single-plant yield. The number of pods per plant (5.83%), seed unit weight (5.66%), pod length (4.53%), and seed-to-pod ratio (4.20%) were identified as variables providing moderate contributions. The fact that the cumulative contribution of the first nine variables reached 90.43% indicates that a significant portion of the model predictions is driven by a limited number of yield components.

In contrast, the contribution of phenological and general morphological traits to model predictions was more limited. Maturation and flowering period contributed only 1.13% and 1.07%, respectively. Plot yield, number of branches, plant height, and first pod height also had low SHAP importance values. The lowest contributions were observed for vegetation duration (0.25%) and 100-seed weight (0.12%). In particular, the low contribution of 100-seed weight indicates that seed size alone is insufficient to explain single-plant yield in the current dataset.

Overall, the SHAP analysis revealed that the TabPFN model's predictions are primarily based on yield components related to pod and seed development. Pod weight, number of seeds per plant, and the pod length-yield index emerged as the model's key determinants, while biological yield and pod filling rate served as additional variables supporting this predictive framework. In contrast, the lower contribution of phenological and general morphological traits suggests that, in the current dataset, single-plant yield prediction is primarily shaped by direct reproductive organs, seed formation, and biomass accumulation.

The SHAP summary plot was used to assess the direction and magnitude of the effect of each variable on the single-plant yield prediction in the TabPFN model. As shown in Figure 4 below, pod weight, number of seeds per plant, and pod length-yield index emerged as the variables with the widest range of SHAP values. This indicates that these traits play a decisive role in the model predictions. The fact that high values for these variables are predominantly in the positive SHAP region, while low values are in the negative SHAP region, suggests that increases in pod weight, seed number, and pod density tend to increase the predicted single-plant yield.

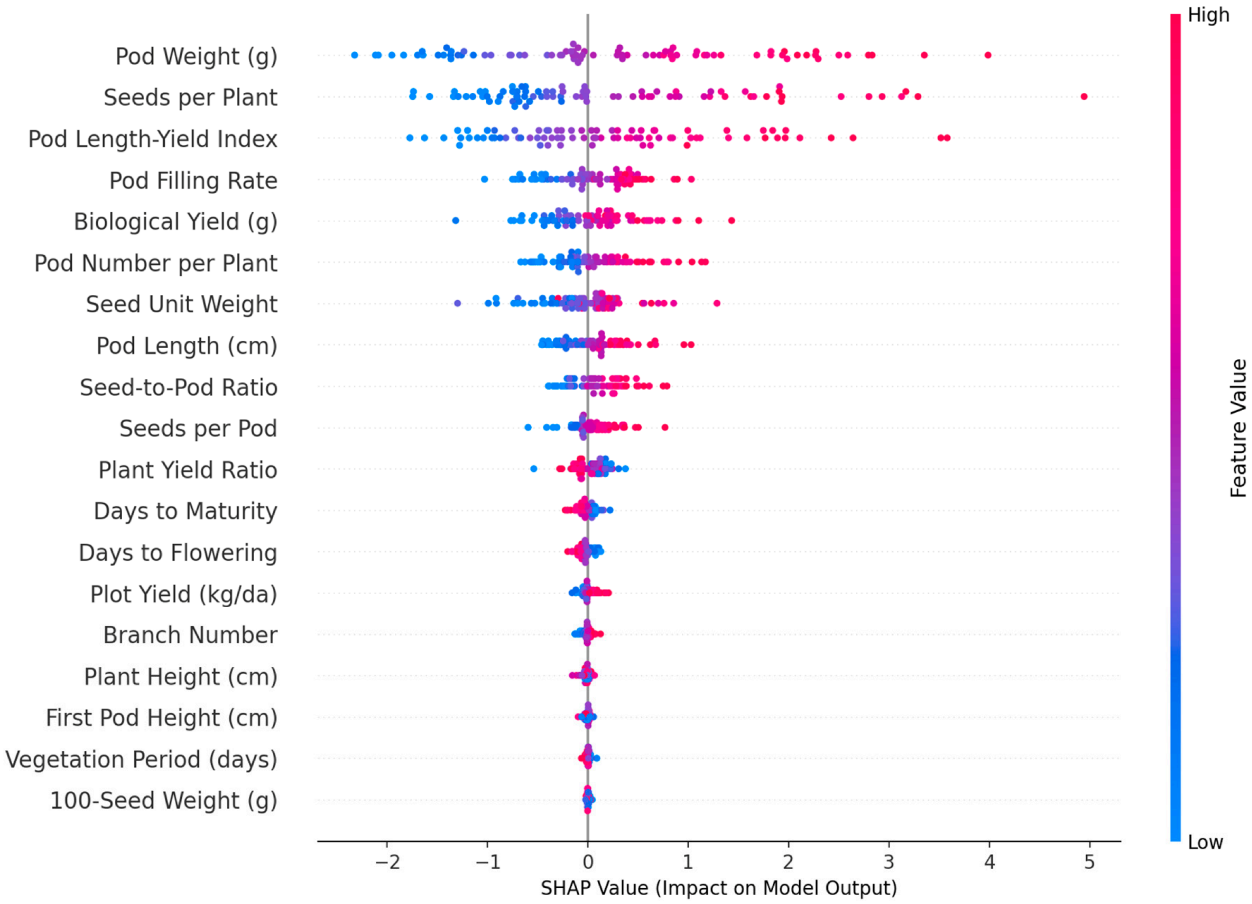

Figure 4. SHAP summary plot of the TabPFN model for single-plant yield estimation.

Pod filling rate, biological yield, number of pods per plant, and seed weight were also among the variables contributing to the model predictions; however, the SHAP value ranges for these traits were more limited compared to the first three variables. In contrast, phenological traits, general morphological traits, and 100-seed weight were concentrated around SHAP values close to zero. This finding indicates that, in the current dataset, the TabPFN model relies more heavily on traits directly related to pods, seeds, and biomass for single-plant yield prediction, while phenological and general morphological traits provide a more limited contribution.

Mean absolute SHAP values were evaluated to summarize the overall importance levels of the variables in the TabPFN model. As shown in Figure 5, pod weight, number of seeds per plant, and the pod length-yield index exhibited significantly higher importance values compared to other variables. In particular, the marked decline in SHAP importance values observed after the third variable indicates that model predictions are largely driven by a limited number of strong yield components.

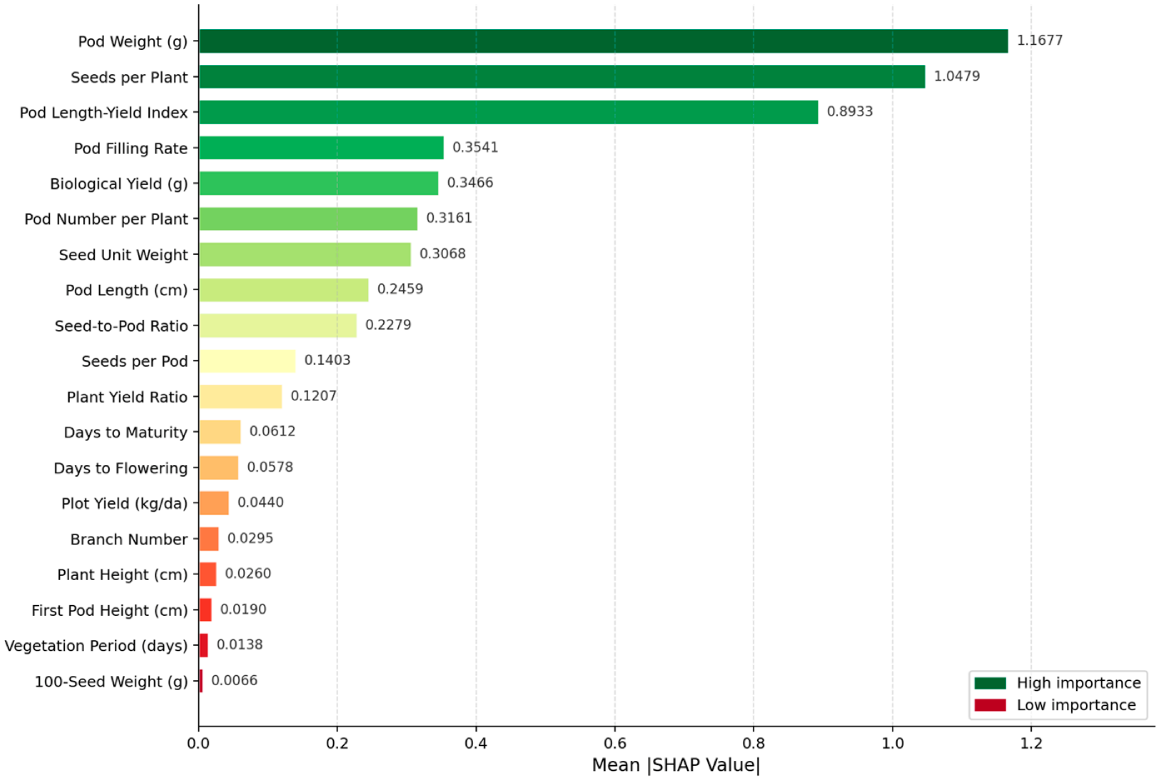

Figure 5. Mean absolute SHAP importance values of input variables in the TabPFN model.

These results indicate that the TabPFN model primarily relies on traits directly related to pod and seed development for single-plant yield prediction. Pod filling rate and biological yield also contributed to the model predictions, but their effects were more limited compared to the top three variables. Overall, the SHAP importance ranking indicates that pod weight, number of seeds per plant, and the pod length-yield index are the primary determinants in single-plant yield prediction, while phenological and general morphological traits contribute to a lesser extent.

Overall, the SHAP results indicate that the TabPFN model relies not on random or biologically weak variables in its single-plant yield prediction, but rather on pod, seed, and biomass components directly related to yield formation. This indicates that the model not only provides high estimation accuracy but also offers an agronomically interpretable and biologically consistent basis for trait prioritization.

To deepen the biological interpretation, relationships among phenological, morphological, biomass-related, and reproductive traits were considered jointly using the global SHAP rankings, Pearson correlations, permutation importance, LOCO results, and established agronomic knowledge. Formal SHAP interaction values and SHAP dependence plots were not computed for the TabPFN model; accordingly, the present analysis does not quantify interaction effects or establish causal relationships. The results instead indicate that correlated harvest-stage variables can share or substitute predictive information, so

low marginal importance should not be interpreted as evidence that a phenological or morphological trait lacks biological relevance.

4.4. Permutation Importance Analysis

Permutation importance analysis was performed on the trained TabPFN model using the held-out test set. A larger decrease in indicates that the corresponding variable contributes more strongly to the predictive performance of the trained model. The results are presented in Table 5 below.

Table 5. Permutation importance of input variables in the TabPFN model based on the mean decrease in test-set R^2.

Rank	Feature	Mean R^2 Drop	Std R^2 Drop	Interpretation
1	Pod Weight (g)	0.1340	±0.0118	Highly Important
2	Seeds per Plant	0.1185	±0.0239	Highly Important
3	Pod Length-Yield Index	0.0930	±0.0262	Highly Important
4	Seed Unit Weight	0.0479	±0.0130	Highly Important
5	Biological Yield (g)	0.0446	±0.0214	Highly Important
6	Pod Filling Rate	0.0145	±0.0051	Important
7	Seeds per Pod	0.0093	±0.0019	Important
8	Pod Length (cm)	0.0077	±0.0047	Important
9	Pod Number per Plant	0.0050	±0.0044	Minor Contribution
10	Seed-to-Pod Ratio	0.0033	±0.0023	Minor Contribution
11	Days to Maturity	0.0028	±0.0012	Minor Contribution
12	Plot Yield (kg da^{-1})	0.0012	±0.0009	Minor Contribution
13	Vegetation Period (days)	0.0008	±0.0002	Minor Contribution
14	Plant Height (cm)	0.0005	±0.0017	Minor Contribution
15	100-Seed Weight (g)	−0.0000	±0.0002	Negligible/Noisy
16	Plant Yield Ratio	−0.0001	±0.0029	Negligible/Noisy
17	First Pod Height (cm)	−0.0004	±0.0007	Negligible/Noisy
18	Branch Number	−0.0006	±0.0004	Negligible/Noisy
19	Days to Flowering	−0.0010	±0.0012	Negligible/Noisy

Importance categories were assigned descriptively based on the relative magnitude of the mean R^2 drop and were used only to facilitate interpretation.

The results showed that pod weight, number of seeds per plant, and pod length-yield index produced the largest decreases in R^2 when permuted, indicating that these variables contributed most strongly to the predictive performance of the trained TabPFN model. Seed unit weight and biological yield also showed substantial permutation effects, confirming the importance of seed mass-related and biomass-related information in the prediction process. In contrast, several phenological and general morphological traits, such as days to flowering, branch number, and first pod height, produced very small or negative R^2 changes. Such near-zero or negative values do not necessarily imply biological irrelevance; rather, they indicate that the variable provided limited additional predictive information within the current feature set, possibly because of redundancy, compensability, or noise. Overall, the permutation importance results support the conclusion that traits directly related to pod development, seed formation, and biomass accumulation are the main drivers of single-plant yield estimation in the trained TabPFN model.

4.5. Leave-One-Covariate-Out (LOCO) Feature Contribution Analysis

The Leave-One-Covariate-Out (LOCO) analysis was conducted to evaluate the unique predictive contribution of each input variable when the model was allowed to relearn without that variable. A positive R^2 drop indicates that the removed variable provided unique predictive information that could not be fully compensated for by the remaining variables. In contrast, a negative R^2 drop indicates that model performance improved after

the variable was removed, suggesting that the variable may be redundant, compensable, or a potential source of noise within the current feature set. The results are presented in Table 6.

Table 6. LOCO feature contribution analysis of the TabPFN model based on the change in test-set R^2 after removing each input variable.

Rank	Feature	R^2 Without	RMSE Without	R^2 Drop	Interpretation
1	Pod Length-Yield Index	0.8669	1.9707	+0.0077	Important
2	Seed Unit Weight	0.8695	1.9520	+0.0052	Important
3	Seeds per Pod	0.8712	1.9390	+0.0034	Minor
4	Seed-to-Pod Ratio	0.8719	1.9335	+0.0027	Minor
5	Pod Length (cm)	0.8724	1.9301	+0.0022	Minor
6	Vegetation Period (days)	0.8728	1.9272	+0.0019	Minor
7	Plot Yield (kg da^{-1})	0.8736	1.9205	+0.0010	Minor
8	Days to Maturity	0.8740	1.9179	+0.0006	Minor
9	Plant Yield Ratio	0.8743	1.9152	+0.0003	Minor
10	Branch Number	0.8749	1.9108	−0.0003	Minor
11	Biological Yield (g plant^{-1})	0.8751	1.9096	−0.0005	Minor
12	100-Seed Weight (g)	0.8753	1.9079	−0.0007	Minor
13	Plant Height (cm)	0.8754	1.9072	−0.0008	Minor
14	Pod Number per Plant	0.8759	1.9036	−0.0012	Minor
15	First Pod Height (cm)	0.8759	1.9030	−0.0013	Minor
16	Days to Flowering	0.8766	1.8981	−0.0019	Minor
17	Pod Weight (g)	0.8766	1.8980	−0.0020	Minor
18	Pod Filling Rate	0.8767	1.8971	−0.0021	Minor
19	Seeds per Plant	0.8804	1.8682	−0.0058	Redundant/ Compensable

Importance categories were assigned descriptively based on the magnitude and direction of the R^2 drop, and were used only to facilitate interpretation.

The results showed that pod length-yield index and seed unit weight produced the largest positive R^2 drops in the LOCO analysis, indicating that these variables provided relatively unique and harder-to-replace predictive information for the TabPFN model. Smaller positive R^2 drops were observed for seeds per pod, seed-to-pod ratio, pod length, vegetation period, plot yield, days to maturity, and plant yield ratio, suggesting limited but non-negligible unique contributions. In contrast, removing some variables did not reduce model performance and even produced very small improvements. In particular, the performance increases observed after removing seeds per plant, pod weight, and pod filling rate should not be interpreted as evidence that these variables are biologically unimportant. Rather, this pattern suggests that their predictive information may overlap with other yield-related traits and can be compensated for by the remaining variables when the model is retrained. While SHAP and permutation importance analyses evaluate the contribution or sensitivity of variables within the trained full model, LOCO evaluates the model's ability to relearn after a variable is completely removed. Therefore, the LOCO results should be interpreted as evidence of unique, non-redundant predictive contribution rather than as a direct ranking of biological importance.

4.6. Statistical Comparison Analysis

All 15 pairwise comparisons among the six models are reported in Table 7.

The Wilcoxon signed-rank analysis compared the two models' absolute errors within each of the same 80 held-out plot-level observations. This pairing accounts for observation-to-observation differences in prediction difficulty, but it does not make absolute error independent of yield magnitude. At the unadjusted 0.05 level, TabPFN produced lower absolute errors than SVR (raw $p = 0.0020$), HistGradientBoosting ($p = 0.0109$), Random Forest

(p = 0.0221), and XGBoost (p = 0.0487). After Holm correction across all 15 comparisons, only the TabPFN–SVR comparison remained significant (adjusted p = 0.0301).

Table 7. Pairwise comparison of model absolute prediction errors using the Wilcoxon signed-rank test with Holm correction.

Model 1	Model 2	Wilcoxon W	Raw p	Holm-Adjusted p	Result	Effect Size (r)	Model with Lower Absolute Error
TabPFN	**SVR**	976.0	0.0020	0.0301	Significant	0.345	**TabPFN**
TabPFN	**HGB**	1089.0	0.0109	0.1522	Not significant	0.285	**TabPFN**
TabPFN	**RandomForest**	1143.0	0.0221	0.2879	Not significant	0.256	**TabPFN**
XGBoost	**HGB**	1151.0	0.0245	0.2938	Not significant	0.251	**XGBoost**
TabPFN	**XGBoost**	1209.0	0.0487	0.5356	Not significant	0.220	**TabPFN**
TabPFN	**ExtraTrees**	1214.0	0.0515	0.5356	Not significant	0.218	**TabPFN**
ExtraTrees	**HGB**	1306.0	0.1321	1.0000	Not significant	0.168	**ExtraTrees**
RandomForest	**HGB**	1347.0	0.1904	1.0000	Not significant	0.146	**RandomForest**
RandomForest	**ExtraTrees**	1448.0	0.4094	1.0000	Not significant	0.092	**ExtraTrees**
SVR	**XGBoost**	1468.0	0.4660	1.0000	Not significant	0.082	**XGBoost**
SVR	**ExtraTrees**	1492.0	0.5393	1.0000	Not significant	0.069	**ExtraTrees**
XGBoost	**RandomForest**	1500.0	0.5649	1.0000	Not significant	0.064	**XGBoost**
SVR	**RandomForest**	1555.0	0.7552	1.0000	Not significant	0.035	**RandomForest**
SVR	**HGB**	1558.0	0.7662	1.0000	Not significant	0.033	**SVR**
XGBoost	**ExtraTrees**	1585.0	0.8667	1.0000	Not significant	0.019	**XGBoost**

Note: reference split, random_state = 42. Two-tailed Wilcoxon signed-rank tests were applied to paired absolute errors for the same 80 held-out observations; Holm correction was applied across 15 comparisons at $\alpha = 0.05$. Effect size was calculated as $r = |Z| / \sqrt{N}$. Pairs are ordered by increasing raw p-value.

The effect sizes for comparisons between TabPFN and the five baselines ranged from r = 0.218 to 0.345, corresponding descriptively to small-to-moderate effects under conventional r benchmarks. With 80 held-out observations and correction across 15 tests, these results provide limited multiplicity-adjusted evidence for most pairwise differences; no formal power analysis was conducted. Among the five tuned baselines, effect sizes did not exceed r = 0.251 and all Holm-adjusted p-values were at least 0.2938. Thus, no statistically detectable pairwise differences among the baselines were identified in this dataset, but this should not be interpreted as evidence of model equivalence. For non-significant comparisons, the 'Model with lower absolute error' column indicates only the numerical direction of the observed difference and should not be treated as statistical evidence of superiority. The Holm-adjusted p-value matrix for all pairwise Wilcoxon signed-rank comparisons is presented in Figure 6.

The Friedman–Nemenyi analysis complemented the pairwise Wilcoxon–Holm comparisons by evaluating all six models jointly on the same 80 held-out plot-level observations. The mean-rank relationships among the models are illustrated in Figure 7.

The corresponding Friedman omnibus test and Nemenyi post-hoc comparison results are summarized in Table 8.

For multiple models evaluated on the same observations, the Friedman test first assesses the omnibus null hypothesis of equal rank performance; post-hoc Nemenyi comparisons are then interpreted after rejection of that hypothesis. Here, the Friedman test rejected the omnibus null ($\chi^2(5) = 16.750$, $p = 5.00 \times 10^{-3}$; Iman–Davenport $F(5, 395) = 3.453$, $p = 4.58 \times 10^{-3}$). TabPFN had the lowest mean rank (2.838) and differed significantly from HistGradientBoosting (Nemenyi p = 0.0023) and SVR (p = 0.0218), whereas its differences from Random Forest, Extra Trees, and XGBoost did not exceed the critical difference of 0.843. Under the equal-performance null, the expected rank for each of six models is 3.5; TabPFN was below this reference value, whereas HistGradientBoosting (3.950) was above it. Kendall's W = 0.042 indicates weak concordance and a small omnibus effect: model

rankings varied substantially across held-out observations, but this statistic alone does not identify distinct plant subgroups or the mechanisms underlying that variation.

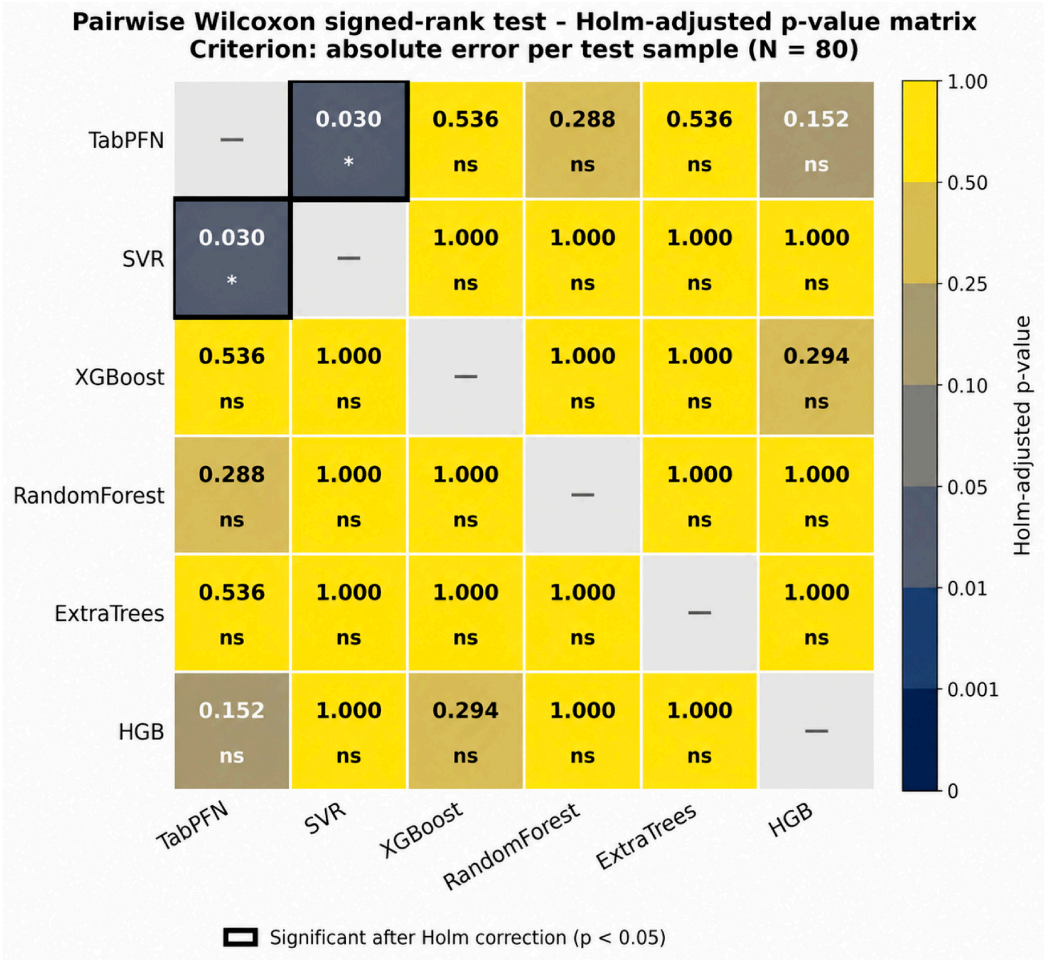

Figure 6. Holm-adjusted *p*-value matrix for pairwise Wilcoxon signed-rank tests on the reference held-out partition (N = 80). Models are ordered by descending test R^2. Black borders and significance codes provide non-colour indicators of statistical significance. (* $p < 0.05$).

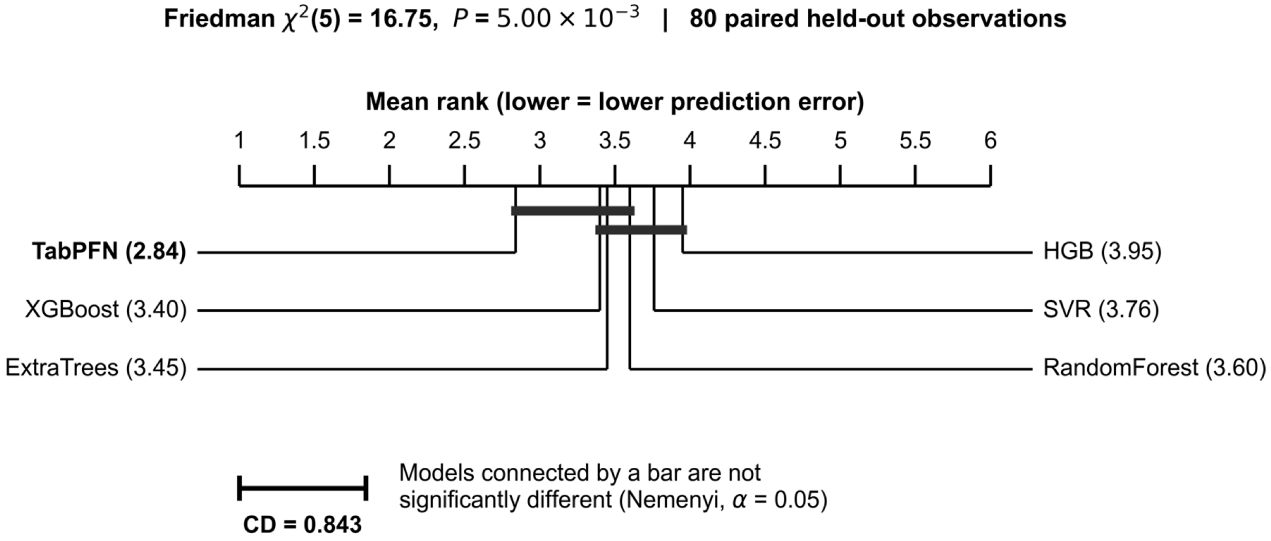

Figure 7. Nemenyi critical-difference diagram for the six models on the reference held-out partition. Models are positioned by mean rank (lower is better); models joined by a horizontal bar are not significantly different at $\alpha = 0.05$ (CD = 0.843).

Table 8. Friedman omnibus test and Nemenyi post-hoc comparisons on the reference held-out partition (N = 80). Friedman $\chi^2(5) = 16.750$, $p = 5.00 \times 10^{-3}$; Iman–Davenport F(5, 395) = 3.453, $p = 4.58 \times 10^{-3}$; Kendall's W = 0.042; Nemenyi CD = 0.843.

Model 1	Model 2	Mean Rank 1	Mean Rank 2	Rank Difference	>CD	Nemenyi *p*	Result
TabPFN	**HGB**	2.838	3.950	1.113	Yes	0.0023	Significant
TabPFN	**SVR**	2.838	3.762	0.925	Yes	0.0218	Significant
TabPFN	**RandomForest**	2.838	3.600	0.763	No	0.1027	Not significant
TabPFN	**ExtraTrees**	2.838	3.450	0.613	No	0.3028	Not significant
TabPFN	**XGBoost**	2.838	3.400	0.562	No	0.4010	Not significant
XGBoost	**HGB**	3.400	3.950	0.550	No	0.4275	Not significant
ExtraTrees	**HGB**	3.450	3.950	0.500	No	0.5382	Not significant
SVR	**XGBoost**	3.762	3.400	0.363	No	0.8245	Not significant
RandomForest	**HGB**	3.600	3.950	0.350	No	0.8451	Not significant
SVR	**ExtraTrees**	3.762	3.450	0.312	No	0.8985	Not significant
XGBoost	**RandomForest**	3.400	3.600	0.200	No	0.9846	Not significant
SVR	**HGB**	3.762	3.950	0.188	No	0.9885	Not significant
SVR	**RandomForest**	3.762	3.600	0.163	No	0.9941	Not significant
RandomForest	**ExtraTrees**	3.600	3.450	0.150	No	0.9959	Not significant
XGBoost	**ExtraTrees**	3.400	3.450	0.050	No	1.0000	Not significant

Note: within each held-out observation, the six models were ranked by absolute error, with rank 1 assigned to the smallest error. A pair was considered significantly different when its mean-rank difference exceeded CD = 0.843.

The significant TabPFN–HistGradientBoosting result from the Nemenyi test, compared with the non-significant Wilcoxon–Holm result for the same pair in Table 7 (adjusted $p = 0.1522$), is not a computational contradiction. Both procedures control family-wise error across the 15 pairwise comparisons, but they use different statistics. Nemenyi compares differences in mean ranks within the joint six-model ranking structure and applies a common critical difference, whereas each Wilcoxon signed-rank test evaluates the paired absolute-error differences for one model pair before Holm correction is applied to the 15 *p*-values. The procedures can therefore yield different decisions for borderline pairs. Agreement across them provides stronger evidence; disagreement indicates that the pairwise conclusion is sensitive to the inferential procedure and should be interpreted cautiously.

4.7. Ablation Study

The ablation analysis compared three predictor configurations across the same 10 repeated 80/20 splits: S1 contained the 13 directly measured traits; S2 contained all 19 measured and derived predictors; and S3 removed pod weight, biological yield, plot yield, and all engineered variables computed from these quantities, leaving 12 predictors. The S3 label is therefore restricted to the specified target-proximal variables rather than implying removal of every harvest-stage yield component.

Table 9 addresses two complementary ablation questions: whether the six derived predictors improve performance (S1 versus S2), and how strongly the models depend on the specified target-proximal yield variables (S2 versus S3).

Adding the derived predictors produced a statistically detectable S1–S2 improvement only for TabPFN: mean held-out test R^2 increased from 0.8552 ± 0.0708 to 0.8614 ± 0.0691 (mean $\Delta R^2 = +0.0062$), with gains in nine of the 10 splits (unadjusted paired Wilcoxon $p = 0.0195$). The S1–S2 differences were not significant for the five tuned baselines; the mean change was negative for Random Forest (-0.0021) and HistGradientBoosting (-0.0051), whereas Extra Trees, XGBoost, and SVR showed positive but non-significant mean changes. Because the derived variables are deterministic transformations of measured traits, they do not add independent measurements; their modest benefit can instead reflect a representation that a particular model can use more efficiently in this finite-sample

setting. The evidence therefore supports a modest, model-dependent benefit rather than a general advantage of feature engineering across all six models.

Table 9. Repeated-split ablation analysis over predictor configurations. S1: 13 measured traits; S2: 19 measured and derived predictors; S3: 12 predictors after removing the specified target-proximal yield variables and their derivatives.

Model	S1—Original Only (13) Test R^2	S2—Original + Derived (19) Test R^2	S3—Specified Target-Proximal Variables Removed (12) Test R^2	S1 → S2 ΔR^2 (p)	S2 → S3 ΔR^2 (p)
TabPFN	0.8552 ± 0.0708	0.8614 ± 0.0691	0.6737 ± 0.0899	+0.0062 (0.0195)	−0.1878 (0.0020)
ExtraTrees	0.8455 ± 0.0621	0.8483 ± 0.0644	0.6576 ± 0.0672	+0.0028 (0.3750)	−0.1908 (0.0020)
XGBoost	0.8272 ± 0.0805	0.8450 ± 0.0623	0.6414 ± 0.0693	+0.0178 (0.1055)	−0.2036 (0.0020)
RandomForest	0.8465 ± 0.0608	0.8443 ± 0.0660	0.6459 ± 0.0762	−0.0021 (0.5566)	−0.1984 (0.0020)
HGB	0.8444 ± 0.0594	0.8393 ± 0.0565	0.6248 ± 0.0733	−0.0051 (0.2754)	−0.2146 (0.0020)
SVR	0.7956 ± 0.1419	0.8255 ± 0.0967	0.6633 ± 0.0858	+0.0299 (0.1934)	−0.1621 (0.0020)

Note: values are mean ± SD of held-out test R^2 over 10 repeated 80/20 splits. S3 excludes pod weight, biological yield, and plot yield together with pod filling rate, seed unit weight, pod length–yield index, and plant yield ratio. ΔR^2 is the mean paired difference across splits; unadjusted two-sided Wilcoxon p-values are shown in parentheses. Because repeated splits overlap, these p-values are interpreted as robustness diagnostics.

Removing the specified target-proximal variables reduced mean test R^2 for every model by 0.1621–0.2146, and all six models declined in every one of the 10 splits (unadjusted paired Wilcoxon $p = 0.0020$ for each model). For TabPFN, mean test R^2 decreased from 0.8614 ± 0.0691 to 0.6737 ± 0.0899. This result quantifies the framework's reliance on harvest-stage, target-proximal predictors and supports its interpretation as a harvest-time trait-estimation and trait-prioritization tool rather than an early-season yield-forecasting system. TabPFN retained the highest mean S3 R^2 (0.6737), narrowly ahead of SVR (0.6633), but ranked first in only five of the 10 S3 splits. Thus, its numerical competitiveness persisted after the specified variables were removed, although this restricted-predictor result does not establish categorical superiority or independence from all yield-related information.

4.8. Robustness Across Repeated Data Splits

Table 10 evaluates whether the model ranking observed on the reference split persisted across 10 repeated random 80/20 partitions of the same 398 plot-level observations.

Table 10. Robustness of model performance across 10 repeated 80/20 splits (seeds 0–9) using the S2 predictor set (19 predictors).

Model	Test R^2 (mean ± SD)	Test RMSE (g plant^{-1})	Test MAE (g plant^{-1})	Test MAPE (%)	Times Ranked Best of 10	Δ R^2 vs TabPFN	TabPFN Better in n of 10	Paired Wilcoxon p
TabPFN	0.8614 ± 0.0691	1.920 ± 0.497	1.056 ± 0.159	8.27 ± 0.94	8	—	—	—
ExtraTrees	0.8483 ± 0.0644	2.033 ± 0.437	1.163 ± 0.130	9.66 ± 1.84	0	0.0131	8	0.0195
XGBoost	0.8450 ± 0.0623	2.061 ± 0.412	1.201 ± 0.113	9.99 ± 2.23	0	0.0164	8	0.0195
RandomForest	0.8443 ± 0.0660	2.063 ± 0.444	1.209 ± 0.135	10.07 ± 2.14	1	0.0171	8	0.0273
HGB	0.8393 ± 0.0565	2.108 ± 0.366	1.272 ± 0.112	10.72 ± 2.69	1	0.0221	8	0.0273
SVR	0.8255 ± 0.0967	2.147 ± 0.587	1.191 ± 0.190	9.37 ± 0.94	0	0.0360	10	0.0020

Note: baseline hyperparameters were optimized once on the reference split (seed = 42; 100 Optuna trials) and held fixed across the 10 repeated splits; TabPFN was refitted with n_estimators = 4 and without per-split retuning. Two-sided paired Wilcoxon tests compare splitwise R^2 differences between TabPFN and each baseline, with Holm correction across the five comparisons. Because repeated splits reuse observations, these tests are interpreted as robustness diagnostics and do not constitute independent external validation.

TabPFN achieved the highest test R^2 in eight of the 10 splits and exceeded each tuned baseline in at least eight splits. Its mean R^2 advantages ranged from 0.0131 to 0.0360, whereas the within-model standard deviations across splits ranged from 0.0565 to 0.0967.

Thus, variation associated with the choice of split was larger than the mean differences between models. The recurrent within-split advantage supports the stability of TabPFN's numerical ranking, but the overlap among repeated splits and the small mean gaps do not establish categorical superiority. Paired splitwise comparisons and win counts preserve the within-split comparison structure, although both remain summaries of resampling the same dataset. SVR showed the largest R^2 standard deviation (0.0967) and the largest mean gap from TabPFN (0.0360), indicating that its estimated performance was the most sensitive to the selected split; the table alone does not identify the observations or mechanisms responsible for that variability.

All five unadjusted Wilcoxon *p*-values were below 0.05, but after Holm correction only the TabPFN–SVR comparison remained significant (adjusted $p = 0.0100$). With only 10 paired splits, exact Wilcoxon *p*-values are discrete: when all 10 non-zero differences have the same sign, the smallest attainable two-sided *p*-value is 0.001953 (reported as 0.0020). However, an eight-of−10 win count does not uniquely determine the Wilcoxon *p*-value because the test also uses the ranks of the absolute paired differences. The non-significant adjusted results for the other baselines therefore cannot be attributed to limited power alone; sample size, effect magnitude, the rank pattern of paired differences, multiplicity correction, and dependence among overlapping splits all contribute. Overall, the repeated-split analysis provides the strongest comparative evidence against SVR and more cautious evidence of a numerical advantage over the remaining tuned baselines.

5. Discussion

In this study, six regression models were evaluated using directly measured agronomic traits and study-specific derived variables to estimate plot-mean single-plant yield in faba bean. On the reference held-out split, TabPFN produced the best values for all four test metrics. However, once each baseline had been optimized with 100 Optuna trials, the between-model differences were small. Holm-corrected Wilcoxon tests confirmed a significant difference only between TabPFN and SVR, whereas the Friedman–Nemenyi analysis distinguished TabPFN from SVR and HistGradientBoosting. TabPFN should therefore be described as numerically strongest and highly competitive in this dataset, rather than as uniformly superior to every tuned comparator.

The repeated-split analysis was used to examine whether the reference result depended strongly on one train/test partition. TabPFN ranked first in eight of 10 splits and produced a higher R^2 than each baseline in at least eight splits, although only the SVR comparison remained significant after correcting the five repeated-split tests. The standard deviation of test R^2 across splits was approximately 0.06–0.10, exceeding most between-model mean differences. Consequently, the repeated analysis supports the stability of TabPFN's numerical ranking but also reinforces the need for cautious claims about small model-to-model differences.

Computational cost provides a separate practical distinction. The five conventional baselines required approximately 1814 s of 100-trial hyperparameter search on the reported workstation, whereas TabPFN did not undergo that Optuna search; only n_estimators was selected from four prespecified values. TabPFN completed reference-split fitting and prediction in approximately 0.51 s. This comparison does not establish hardware-independent speed superiority because TabPFN used the GPU and the baselines used the CPU; it demonstrates that the pretrained model avoided the extensive Bayesian search procedure required by the tuned baselines.

The explainability analyses indicated that TabPFN relied most strongly on pod weight, number of seeds per plant, pod length–yield index, seed unit weight, and biological yield. These quantities are model-attribution results and should not be interpreted as causal effects.

Their emphasis on reproductive and biomass-related traits is consistent with agronomic work on faba bean yield components [2,7] and with explainable or ANN-based analyses in grain legumes that also identified seed- and reproductive-trait information as useful for yield estimation [27,31]. This agreement supports the agronomic plausibility of the ranking, but it does not independently validate the biological mechanisms because several predictors are measured at harvest and are strongly correlated or algebraically related.

Phenological and morphological traits had lower marginal importance in the global rankings. It is agronomically plausible that such traits influence biomass accumulation, pod development, and seed filling indirectly [2,7]; however, the present analysis did not estimate causal pathways or formal feature-interaction effects. The lower marginal attribution of these variables may also reflect the presence of more target-proximal harvest-stage predictors. The indirect relationships discussed here should therefore be treated as biologically informed hypotheses rather than effects demonstrated by the explainability analysis.

The explainability results must be interpreted in light of the strong dependence structure within the predictor set. Because several traits are biologically coupled or algebraically linked, SHAP, permutation importance, and LOCO quantify different aspects of model reliance and are affected differently by feature dependence. SHAP partitions predictive attribution among correlated traits, which can dilute the apparent importance of individual variables within a dependent group. Permutation importance breaks one variable's association with all others and can therefore generate biologically implausible combinations for strongly correlated features. LOCO measures the non-redundant contribution of a variable after the model is refitted; a correlated substitute can compensate when one variable is removed. This dependence helps explain why pod weight and the pod length–yield index exchange positions across rankings and why the perfectly collinear phenological triplet cannot be assigned separate independent contributions. The rankings should therefore not be viewed as a strict biological hierarchy. Agreement among the methods is better interpreted as consistent evidence that the model relies on groups of mutually dependent variables associated with biomass accumulation and pod filling.

The inclusion of harvest-stage predictors closely related to final yield (e.g., pod weight, biological yield, plot yield, and their engineered derivatives) inherently defines the scope of this framework. As explicitly quantified by the ablation study results, removing direct yield components causes a large drop in accuracy for every model. This confirms that the current framework functions fundamentally as a harvest-time trait-estimation and trait-prioritization tool, rather than an early-season predictive system. Because these traits are determined at physiological maturity, they are not available for early-season forecasting. Consequently, the present model cannot be deployed for real-time crop management or pre-harvest decision-making regarding irrigation and fertilization. The reported model performance reflects the framework's ability to reconstruct and biologically explain yield variation once harvest-stage information is available, not its predictive capacity before harvest. In particular, plot yield acts as a target-proximal variable reflecting yield at a different scale, meaning its predictive contribution stems from shared biological information rather than early availability. From a practical breeding perspective, this model can support post-harvest genotype evaluation. It provides a data-driven basis for developing multi-trait selection criteria and weighting traits associated with high yield, but it does not replace early-generation selection methodologies.

A primary limitation of the current study is that the dataset was derived from a single experimental location during one growing season utilizing an augmented field design. Consequently, the data do not capture the full spectrum of genotype by environment (G×E) interactions that occur across diverse agroecological production environments. The relatively large number of genetically diverse observations (398 plots representing

372 landraces and replicated control cultivars) provided adequate phenotypic variability to evaluate the machine learning framework under uniform field conditions; however, the model should be strictly interpreted as a trait-based yield estimation and prioritization approach validated for this specific experimental environment. It is not currently intended as a universally applicable agronomic forecasting model. Furthermore, regarding statistical independence, the dataset was obtained using an augmented field design. While the 372 local genotypes were evaluated as unreplicated single plots, the control cultivars were repeated across the nine blocks to monitor environmental variation. Consequently, although the control plots represent separate experimental units exposed to varying block-level microenvironments, they share the same genetic identity. While this is a standard and necessary practice in early-generation screening of large genetic pools, observations obtained from repeated control cultivars cannot be regarded as genetically independent in the strict sense. To extend this framework to broader agroecological conditions and assess model generalizability under varying environmental stresses, future research must incorporate additional validation using multi-location, multi-year, and independent external datasets. The repeated-split analysis has two additional limitations. First, the 10 partitions reuse observations and are therefore correlated rather than independent external replications. Second, the baseline hyperparameters were optimized on the reference split and then held fixed across the other nine splits, whereas TabPFN did not undergo an analogous 100-trial Bayesian search, although n_estimators was selected from four prespecified values. This design applies the same fixed configurations to all baselines but may mildly favor TabPFN if a baseline's optimal configuration changes across partitions. A stricter future comparison should combine nested retuning within every split with genotype-grouped or external environmental validation.

6. Conclusions

This study estimated plot-mean single-plant yield in faba bean (*Vicia faba* L.) using directly measured agronomic traits and study-specific derived variables and interpreted model reliance through SHAP, permutation importance, and LOCO analyses. On the reference held-out split, TabPFN produced the best numerical values for all four metrics (R^2 = 0.8746; RMSE = 1.9132 g plant^{-1}). After Holm correction, its paired error difference was significant only relative to SVR, while Nemenyi comparisons also distinguished it from HistGradientBoosting. Across 10 repeated splits, TabPFN had the highest mean test R^2 (0.8614 $\pm$ 0.0691) and ranked first in eight splits. These results support TabPFN as a competitive model in the present dataset, while not establishing categorical superiority over the strongest tuned tree-based baselines. Its clearest operational advantage was that it avoided the 100-trial Optuna search applied to each baseline; only n_estimators was screened over four prespecified values.

Beyond predictive performance, the main contribution of this study is the biological interpretability of the proposed framework. SHAP, permutation importance, and LOCO analyses collectively indicated that pod weight, number of seeds per plant, pod length-yield index, seed unit weight, and biological yield were among the most influential variables associated with single-plant yield estimation. These traits are closely related to pod development, seed formation, pod filling, individual seed mass, and biomass accumulation, supporting the agronomic relevance of the model outputs. Therefore, the proposed framework provides a complementary decision-support approach for genotype evaluation, yield-oriented trait prioritization, and the development of interpretable multi-trait selection strategies in faba bean breeding studies. Furthermore, due to the substantial feature dependence among yield components, practical applications aiming to build multi-trait selection indices should select one representative trait per highly correlated group

rather than relying simultaneously on all mutually dependent members identified by the explainability framework.

Nevertheless, the findings are limited to one location, one growing season, and repeated partitions of the same 398 plot-level observations. The repeated-split ablation showed that removing the specified target-proximal yield variables reduced R^2 for every model in all 10 splits; for TabPFN, mean R^2 decreased from 0.8614 to 0.6737. The framework should therefore be regarded as a harvest-time trait-estimation and trait-prioritization approach rather than an early-stage forecasting system. Future work should evaluate a strictly pre-harvest predictor set, retune baseline models within each resampling loop, group repeated genotypes during splitting, and validate performance across independent years and locations.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/agronomy16171653/s1, Table S1: Descriptive information on the broader source collection of local faba bean populations from provinces and regions in Türkiye; Table S2: Hyperparameter search spaces and selected configurations for the reference S2 analysis; Table S3: Computational cost on the reported workstation (20-thread CPU, NVIDIA RTX 4050 Laptop GPU with 6 GB, Windows 11); Data S1: Plot-level faba bean dataset comprising 398 observations, 19 predictor variables (13 original agronomic traits and six derived features), and the single-plant yield target variable.

Author Contributions: Conceptualization, İ.Y., T.K., H.B. and M.Z.; methodology, İ.Y., T.K., H.B. and M.Z.; software, İ.Y., S.K. and M.Z.; validation, Y.Ç., T.K., İ.Y., H.B., S.K. and M.Z.; formal analysis, Y.Ç., T.K., İ.Y., H.B., S.K. and M.Z.; investigation, Y.Ç., T.K. and H.B.; resources, Y.Ç., T.K. and H.B.; data curation, Y.Ç., T.K.; writing—original draft preparation, İ.Y., T.K., H.B., S.K. and M.Z.; writing—review and editing, Y.Ç., İ.Y., T.K., H.B., S.K. and M.Z.; visualization, İ.Y., T.K., H.B. and M.Z.; supervision, S.K. and M.Z.; project administration, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is provided as Supplementary Data S1. The Python 3.13.0 scripts developed for this study are available at https://github.com/metinzontul/faba-bean-tabpfn-xai (accessed on 18 August 2026).

Acknowledgments: During the preparation of this manuscript, the authors used GPT-5 (OpenAI, San Francisco, CA, USA) and [Gemini 2.5 Pro] (Google LLC, Mountain View, CA, USA) for text refinement, language editing, and structural improvement. The authors reviewed and edited all generated outputs and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Karaköy, T.; Baloch, F.S.; Toklu, F.; Özkan, H. Genome-wide association studies revealed DArTseq loci associated with agronomic traits in Turkish faba bean germplasm. *Genet. Resour. Crop Evol.* **2023**, *71*, 181–198. [CrossRef]
2. Mínguez, M.I.; Rubiales, D. Faba bean. In *Crop Physiology Case Histories for Major Crops*; Sadras, V.O., Calderini, D.F., Eds.; Elsevier: Amsterdam, The Netherlands, 2021; pp. 452–481.
3. Rahate, K.A.; Madhumita, M.; Prabhakar, P.K. Nutritional composition, anti-nutritional factors, pretreatments-cum-processing impact and food formulation potential of faba bean (*Vicia faba* L.): A comprehensive review. *LWT* **2021**, *138*, 110796. [CrossRef]
4. Kumar, T.; Hamwieh, A.; Swain, N.; Sarker, A. Identification and morphological characterization of promising kabuli chickpea genotypes for short-season environment in central India. *J. Genet.* **2021**, *100*, 33. [CrossRef]

5. Keneni, G.; Jarso, M.; Wolabu, T.; Dino, G. Extent and pattern of genetic diversity for morpho-agronomic traits in Ethiopian highland pulse landraces. II. Faba bean (*Vicia faba* L.). *Genet. Resour. Crop Evol.* **2005**, *52*, 551–561. [CrossRef]
6. Karaköy, T.; Baloch, F.S.; Toklu, F.; Özkan, H. Variation for selected morphological and quality-related traits among 178 faba bean landraces collected from Turkey. *Plant Genet. Resour.* **2014**, *12*, 5–13. [CrossRef]
7. Loss, S.P.; Siddique, K.H.M. Adaptation of faba bean (*Vicia faba* L.) to dryland Mediterranean-type environments. I. Seed yield and yield components. *Field Crops Res.* **1997**, *52*, 17–28. [CrossRef]
8. Bakır, H. Evaluating the impact of tuned pre-trained architectures' feature maps on deep learning model performance for tomato disease detection. *Multimed. Tools Appl.* **2024**, *83*, 18147–18168. [CrossRef]
9. Hernández Hernández, G.C.; Gómez Gómez, J.; Jiménez-Cabas, J. Predictive models based on artificial intelligence to estimate crop yield: A literature review. *Agriculture* **2025**, *15*, 2438. [CrossRef]
10. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
11. Albahri, A.S.; Alnoor, A.; Zaidan, A.A.; Albahri, O.S.; Hameed, H.; Zaidan, B.B.; Peh, S.S.; Zain, A.B.; Siraj, S.B.; Masnan, A.H.B.; et al. Hybrid artificial neural network and structural equation modelling techniques: A survey. *Complex Intell. Syst.* **2022**, *8*, 1781–1801. [CrossRef] [PubMed]
12. Kaya, M.; Varışlı, O.; Yılmaz, A.; Acay, Ü.; Ipekeşen, S.; Guevara, L.; Bicer, B.; Budak, C.; Valluru, R. Adaptive capacity of machine learning approaches for leaf area estimation in faba bean under diverse agro-environmental conditions. *J. Agr. Sci.-Tarim Bili.* **2026**, *32*, 439–454. [CrossRef]
13. Ji, Y.; Liu, Z.; Liu, R.; Wang, Z.; Zong, X.; Yang, T. High-throughput phenotypic traits estimation of faba bean based on machine learning and drone-based multimodal data. *Comput. Electron. Agric.* **2024**, *227*, 109584. [CrossRef]
14. Cui, Y.; Ji, Y.; Liu, R.; Li, W.; Liu, Y.; Liu, Z.; Zong, X.; Yang, T. Faba bean (*Vicia faba* L.) yield estimation based on dual-sensor data. *Drones* **2023**, *7*, 378. [CrossRef]
15. Ji, Y.; Chen, Z.; Cheng, Q.; Liu, R.; Li, M.; Yan, X.; Li, G.; Wang, D.; Fu, L.; Ma, Y.; et al. Estimation of plant height and yield based on UAV imagery in faba bean (*Vicia faba* L.). *Plant Methods* **2022**, *18*, 26. [CrossRef] [PubMed]
16. Ji, Y.; Liu, R.; Xiao, Y.; Cui, Y.; Chen, Z.; Zong, X.; Yang, T. Faba bean above-ground biomass and bean yield estimation based on consumer-grade unmanned aerial vehicle RGB images and ensemble learning. *Precis. Agric.* **2023**, *24*, 1439–1460. [CrossRef]
17. Mohammadi, S.; Uhlen, A.K.; Lillemo, M.; Ergon, A.; Shafiee, S. Enhancing phenotyping efficiency in faba bean breeding: Integrating UAV imaging and machine learning. *Precis. Agric.* **2024**, *25*, 1502–1528. [CrossRef]
18. Tang, Y.; Fitzgerald, G.J.; Gupta, D.; Delahunty, A.; Nuttall, J.G.; Walker, C. Predicting faba bean yield and grain quality pre-harvest using chemometric modelling. *Precis. Agric.* **2026**, *27*, 6. [CrossRef]
19. Lippolis, A.; Polo, P.V.; de Sousa, G.; Dechesne, A.; Pouvreau, L.; Trindade, L.M. High-throughput seed quality analysis in faba bean: Leveraging Near-InfraRed spectroscopy (NIRS) data and statistical methods. *Food Chem. X* **2024**, *23*, 101583. [CrossRef] [PubMed]
20. Poyraz, İ.; Akçura, M. Binary classification of faba bean (*Vicia faba* L.) cultivars based on appearances using image processing technique and machine learning algorithms. *Turk. J. Agric. Nat. Sci.* **2025**, *12*, 1084–1108. [CrossRef]
21. Aboukarima, A.; El-Marazky, M.; Elsoury, H.; Zayed, M.; Minyawi, M. Artificial neural network-based method to identify five varieties of Egyptian faba bean according to seed morphological features. *Eng. Agríc.* **2020**, *40*, 791–799. [CrossRef]
22. Torabian, S.; Farhangi-Abriz, S. Integrating machine learning approaches to evaluate faba bean varietal performance under fall and spring plantings. *Smart Agric. Technol.* **2025**, *12*, 101570. [CrossRef]
23. Salehi, F.; Ghazvineh, S.; Amiri, M. Investigating the effect of microwave power on moisture diffusivity and drying kinetics of faba beans: An experimental and modeling study. *J. Food Sci. Technol.* **2025**, *21*, 213–224.
24. Kaul, M.; Hill, R.L.; Walthall, C. Artificial neural networks for corn and soybean yield prediction. *Agric. Syst.* **2005**, *85*, 1–18. [CrossRef]
25. Pant, J.; Pant, R.P.; Singh, M.K.; Singh, D.P.; Pant, H. Analysis of agricultural crop yield prediction using statistical techniques of machine learning. *Mater. Today Proc.* **2021**, *46*, 10922–10926. [CrossRef]
26. Rashid, M.; Bari, B.S.; Yusup, Y.; Kamaruddin, M.A.; Khan, N. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access* **2021**, *9*, 63406–63439. [CrossRef]
27. Li, Y.; Li, R.; Ji, R.; Wu, Y.; Chen, J.; Wu, M.; Yang, J. Research on factors affecting global grain legume yield based on explainable artificial intelligence. *Agriculture* **2024**, *14*, 438. [CrossRef]
28. Hara, P.; Piekutowska, M.; Niedbała, G. Prediction of protein content in pea (*Pisum sativum* L.) seeds using artificial neural networks. *Agriculture* **2022**, *13*, 29. [CrossRef]
29. Rosado, R.D.S.; Cruz, C.D.; Barili, L.D.; de Souza Carneiro, J.E.; Carneiro, P.C.S.; Carneiro, V.Q.; da Silva, J.T.; Nascimento, M. Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars. *Agriculture* **2020**, *10*, 638. [CrossRef]

30. Guo, Y.; Xiang, H.; Li, Z.; Ma, F.; Du, C. Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression. *Agronomy* **2021**, *11*, 282. [CrossRef]
31. Karakoy, T.; Yelmen, I.; Zontul, M.; Yildirim, F. Predicting chickpea yield using artificial neural networks with explainable AI. *Agronomy* **2026**, *16*, 768. [CrossRef]
32. Agboka, K.M.; Tonnang, H.E.Z.; Abdel-Rahman, E.M.; Odindi, J.; Mutanga, O.; Niassy, S. Data-driven artificial intelligence (AI) algorithms for modelling potential maize yield under maize–legume farming systems in East Africa. *Agronomy* **2022**, *12*, 3085. [CrossRef]
33. Shekoofa, A.; Emam, Y.; Shekoufa, N.; Ebrahimi, M.; Ebrahimie, E. Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: A new avenue in intelligent agriculture. *PLoS ONE* **2014**, *9*, e97288. [CrossRef] [PubMed]
34. Dos Santos Silva, P.P.; Sousa, M.; de Oliveira, E.J. Prediction models and selection of agronomic and physiological traits for tolerance to water deficit in cassava. *Euphytica* **2019**, *215*, 73. [CrossRef]
35. Mohammadi Mirik, A.; Parsaeian, M.; Rohani, A.; Lawson, S. Optimizing linseed (*Linum usitatissimum* L.) seed yield through agronomic parameter modeling via artificial neural networks. *Agriculture* **2024**, *14*, 25. [CrossRef]
36. Fuentes, I.; Al-Shammari, D. Field-aware and explainable modelling for early-season crop yield prediction using satellite-derived phenology. *Remote Sens.* **2026**, *18*, 890. [CrossRef]
37. Dey, A.; Saha, A.; Sarkar, S.; Mondal, A.; Mitra, P. XAI-driven feature reduction for improved agricultural yield prediction. *Multimed. Tools Appl.* **2026**, *85*, 182. [CrossRef]
38. Elbeltagi, A.; Srivastava, A.; Cao, X.; El Bilali, A.; Raza, A.; Khadke, L.; Salem, A. An interpretable machine learning approach based on SHAP, Sobol and LIME values for precise estimation of daily soybean crop coefficients. *Sci. Rep.* **2025**, *15*, 36594. [CrossRef] [PubMed]
39. Jabed, M.A.; Murad, M.A.A. Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon* **2024**, *10*, e40836. [CrossRef] [PubMed]
40. Dangi, A.; Rathore, D.K. A LightGBM-SHAP framework for accurate and interpretable crop yield prediction. In Proceedings of the 2025 IEEE International Conference on Recent Advances in Computing and Systems (REACS), Gwalior, India, 19–20 December 2025.
41. Kassem, M.A. Harnessing artificial intelligence and machine learning for identifying quantitative trait loci (QTL) associated with seed quality traits in crops. *Plants* **2025**, *14*, 1727. [CrossRef]
42. Zhang, Z.; Li, Y.; Xie, L.; Li, S.; Feng, H.; Siddique, K.H.M.; Lin, G. Game analysis of future rice yield changes in China based on explainable machine-learning and planting date optimization. *Field Crops Res.* **2024**, *317*, 109557. [CrossRef]
43. International Board for Plant Genetic Resources (IBPGR); International Center for Agricultural Research in the Dry Areas (ICARDA). *Faba Bean Descriptors*; IBPGR Secretariat: Rome, Italy, 1985.
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
45. Scikit-Learn Developers. Sklearn.Preprocessing.QuantileTransformer. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html (accessed on 22 June 2026).
46. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019.
47. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
48. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
49. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.