

Wavelength heritability, collinearity and bias influences prediction accuracy in phenomic selection on rapeseed hybrids

Andris Justin Finkbeiner

`andris.finkbeiner@uni-giessen.de`

University of Giessen

Anika Unger

University of Giessen

Andreas Stahl

Julius Kühn-Institut

Benjamin Wittkop

University of Giessen

Rod Snowdon

University of Giessen

Lennard Roscher-Ehrig

University of Giessen

Sven Ernst Weber

University of Giessen

Research Article

Keywords: NIRS, Brassica napus, multicollinearity, molecular proxies, HBLUP, GBLUP, genotype-environment interactions

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9049170/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare no financial interest that is directly related to the work submitted for publication. Non-financial interests: Author Rod Snowdon is Editor in Chief of Theoretical and Applied Genetics.

Wavelength heritability, collinearity and bias influences prediction accuracy in phenomic selection on rapeseed hybrids

Andris J. Finkbeiner^{*1}, Anika Unger¹, Andreas Stahl², Benjamin Wittkop¹, Rod Snowdon¹, Lennard Roscher-Ehrig¹, Sven E. Weber¹

¹Justus-Liebig University Giessen
Department of Plant Breeding
IFZ - Research Center for Biosystems, Land Use and Nutrition
Heinrich-Buff-Ring 26-32
35392 Giessen, Germany

²Julius Kuehn-Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Erwin-Baur-Straße 27, Quedlinburg, Germany

*corresponding author

andris.finkbeiner@uni-giessen.de

<https://orcid.org/0009-0005-6199-5173>

Acknowledgments

The authors thank NPZ Innovation GmbH and Limagrain GmbH for providing data and seed samples from field trials. Furthermore, the authors thank Paul Kimwemwe and Mohammed Mosalam for reading and improving the manuscript and Zsa Zsa Boyny, Nuri Güvencli and Daniel Valle Torres for discussing the methodology. We thank the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics (de.NBI) for providing computational infrastructure. Fig. 1 and Fig. 2 were created with BioRender

Funding

This work was funded by Federal Ministry of Research, Technology and Space with grant 031B1539 and Federal Ministry for Food and Agriculture grant 281B200416, with additional support by DFG grant 469336000 from the German Research Society (DFG) for the International Research Training Group IRTG 2843 Accelerating Crop Genetic Gain

Competing Interests

The authors declare no financial interest that is directly related to the work submitted for publication. Non-financial interests: Author Rod Snowdon is Editor in Chief of Theoretical and Applied Genetics.

Author Contributions

S.E.W., L.R. and A.J.F. contributed to the study conception and design. Primary analysis of field trial and NIRS data was performed by A.S., L.R. and B.W.. The data analysis was performed by A.J.F.. Supervision of the study was done by S.E.W.. A.U., L.R. and S.E.W. provided support in R coding and server infrastructure. The first draft of the manuscript was written by A.J.F. . All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on request.

Abstract

Phenomic selection has emerged as a promising yet controversially discussed alternative to genomic selection. Its appeal lies in the relatively low cost of marker acquisition and the increasing availability of phenotypic data. Crops like rapeseed, in which Near Infrared Spectroscopy (NIRS) is routinely applied for assessment of seed quality, can benefit from phenomic selection without any additional costs as the spectral data is already collected routinely in breeding programs. Despite this potential, several challenges hinder its effective implementation. Among these, the quality of spectral wavelengths is frequently debated as a key factor influencing model performance. Broad-sense wavelength heritability is commonly employed as an indicator of wavelength relevance in NIRS-based prediction models. Here we demonstrate that heritability alone does not fully capture the predictive relevance of individual wavelengths. Owing to the high degree of collinearity among wavelengths, only a small fraction of approximately 1.1% is necessary for accurate trait prediction. Moreover, phenomic selection models often rely on molecular proxies that are correlated with the target trait. In rapeseed, for instance, protein or oil content are correlated to seed yield and lead to a potential bias for such traits, which may ultimately reduce the response to selection and undesired selection pressure for correlated traits. In contrast, we found that other traits like flowering time are lacking such molecular correlations and appear to be unaffected by this issue.

Keywords

NIRS, Brassica napus, multicollinearity, molecular proxies, HBLUP, GBLUP, genotype-environment interactions

Introduction

A major challenge for plant breeders is to estimate the genotypic effect on the plant's performance, as it is the only effect that can be truly impacted by breeding (Piepho et al. 2008; Lorenzana and Bernardo 2009). Prediction of the performance of unobserved

plants based on their genotypic information can save laborious and costly field-trials, while also shortening the timeframe for selection of superior genotypes. The most widely used prediction method, especially in important grain crops, is genomic selection (GS) (Meuwissen et al. 2001; Würschum et al. 2014; Hu et al. 2019; Sweeney et al. 2021). For GS, the cumulative effect of individual markers is estimated to compare genomic estimated breeding values in non-phenotyped individuals. The GS model requires prior training on a dataset that includes both phenotypic information and genotypic data. Afterwards, the marker effects can be applied as predictors on untested genotypes, saving time and costs by pre-selecting these based on the predicted performance (Ou and Liao 2019).

Even though this method has achieved good results in prediction accuracy in crops like wheat (Hu et al. 2019), rapeseed (Würschum et al. 2014) or barley (Sweeney et al. 2021), the costs and required lab work can limit its application in breeding programs with economic constraint. For this reason, there is a great demand for cost-effective, high-throughput alternatives that can achieve comparable prediction results without the high establishment costs associated with development of genotyping platforms. This challenge was addressed in the last few years by utilizing near-infrared spectroscopy (NIRS), where reflectance values at individual wavelengths were used to train models for prediction of complex phenotypic traits (Rincent et al. 2018). This so called phenomic selection (PS) is possible due to the differential absorbance of near-infrared (NIR) light by organic compounds, such as proteins, water or lipids, particularly by C-H, N-H and O-H bonds (Williams et al. 1985; Foley et al. 1998; Davies 2005), which themselves can in some way be associated to the genetic makeup of the plant.

Already in 1962, NIRS was used to measure moisture content in different seeds including cereals like wheat and barley, as well as oil seeds like flaxseed and soybean (Hart et al. 1962). In the following years, NIRS was calibrated on many compounds, such as oil content (OC), starch and protein content (PC) and was used to predict their concentration in grains like rapeseed kernels (Tkachuk 1981). Later, NIRS was used to distinguish between different genotypes of coffee and melon, assuming that different genetics would also cause differences in composition and, therefore, absorption of NIR spectra (Seregély et al. 2004). These findings underpin its ability to transport information about genotypic properties. Subsequently, Rincent et al. (2018) proposed to use NIRS data to compute relationship matrices for the prediction of complex traits, similar to what is done with genomic markers during genomic selection. They investigated wavelength regions associated with polygenic relationships or QTL and their heritability, concluding that these two types of wavelengths are most useful for PS, as they capture both relatedness and linkage disequilibrium. According to their findings, these properties are reflected in wavelength heritability (Rincent et al. 2018).

It must be mentioned that NIRS is impacted by environmental noise. It has been shown that the dependency of NIR absorption ranges from sample temperature and

environmental temperature to lamp aging (Büning-Pfaue 2003; Pojić et al. 2012) as well as the granulometry of the samples (Rincent et al. 2018). For this reason, and to take effects of the experimental design into account, pre-processing of the raw data must be done. Commonly, adjusting of the raw data with a mixed linear model, subsequently smoothing of the data with the Savitsky-Golay (SG) approach, and centering and scaling is conducted (Robert et al. 2022b; Zhu et al. 2022; Dallinger et al. 2023; Braun et al. 2025). The SG smoothing calculates a polynomial regression curve, using the least square approach of a window of the data with given size. When the coefficients are found, they are substituted back into the equation with the central x value for the block. The resulting y value is the best-fitted point at a given x with the least squares approach (Savitzky and Golay 1964). In this approach, the smoothing ability can be controlled by the window size, the polynomial order and different differential orders, which can be used to extract different types of information (Zimmermann and Kohler 2013).

The prediction accuracy for NIRS based prediction models can vary greatly between crops, sample, and trait (Rincent et al. 2018; Zhu et al. 2021; Robert et al. 2022a). Roscher-Ehrig et al. (2024) compared GS models with PS models, demonstrating that PS has comparable and sometimes higher prediction accuracy than GS models. In contrast, several studies have shown that PS often exhibits lower prediction accuracy than GS, especially for monogenic traits, with certain tissues like wood delivering poor predictions (Gebreselassie et al. 2017; Rincent et al. 2018; Brault et al. 2022). However, the economic superiority of NIRS-BLUP models in comparison to different genomic marker-based models is high: Rincent et al. (2018) estimated an expected selection gain for grain yield in a set breeding program budget in the range from -10% to 222%, depending on the tissue and environment. Interestingly, they also showed that the expected gain with NIRS-BLUP models is significantly lower for less complex traits, such as rust resistance (-10-21%) in wheat.

Both, Montesinos-López et al. (2017) and Robert et al. (2022) proposed to use only wavelengths with high heritability for NIRS-based phenomic prediction, since they are likely to have greater influence on the predictive ability than wavelengths with lower heritability. Heritability provides information on the degree to which genotypic variance contributes to overall variance and depends on the variance components of a given model (Schmidt et al. 2019). Genotypic variance, in this case, refers to the genotypic variance for each wavelength. Thus, high genotypic variance for each wavelength allows for the selection of cultivars that exhibit genotype-specific reflection patterns, which can be linked to traits by directly referring to trait-dependent molecular patterns (Robert et al. 2022b). To what extent the wavelength heritability can be used as quality indicator for single wavelength, and the minimum spectral resolution required, has to date not been fully elucidated. Therefore, this study aimed to investigate and optimize heritability thresholds for wavelengths to be included in a prediction model.

In a second step, the number of wavelengths included in the model was reduced by eliminating wavelengths (i) according to heritability, (ii) by increasing the intervals between wavelengths, and (iii) by using a sliding window approach to identify potentially important regions of the wavelength spectrum.

Materials and methods

Trial Setup

The dataset underlying this study, comprising both phenotypic and genotypic data, has been published previously by Roscher-Ehrig et al. (2024). All field trials and data acquisition procedures carried out by Roscher-Ehrig et al. (2024) are detailed in the following section.

Genetic material

The population used consists of 410 oilseed rape test hybrids created by NPZ Innovation GmbH (Holtsee, Germany) and Limagrain GmbH (Peine-Rosenthal, Germany). This population was formed by crossing 251 homozygous genotypes with two different testers that exhibit male sterility to prevent self-fertilization. The two testers are referred to as ‘mothers’, as they served as female parents. In contrast, the other 251 genotypes are referred to as ‘fathers’, as they served as pollinators. The pollinators, in turn, were created using 5 founder lines, all of which were crossed with the same elite line. Therefore, the resulting population is structured into five subfamilies, each containing 45 to 53 different genotypes. Of the 251 pollinators, 159 were crossed with both mothers, while 46 were crossed only with one of the mothers, respectively.

Data collection

The phenotypic data were obtained from field trials in two seasons (2019/2020 and 2020/2021) and 5 locations in Germany, namely Rauischolzhausen (RHH), Lauenau (LAU), Moosburg (MOO), Hohenlieth (HOH) and Rosenthal (ROS). The design was partially replicated. In each location, 20% of the genotypes were replicated once, while 80% were unreplicated. In this way, each genotype was grown in duplicate at one of the 5 locations. Additionally, 3 different check genotypes were used to obtain a more precise adjustment of design effects. These checks were replicated 10 times in each environment. All 251 pollinators have been grown unreplicated in RHH in both years, in a separate trial with the same checks replicated 10 times each.

Traits

The traits considered in this study were seed yield (SY) adjusted for standard water content of 9%, mean plant height (PH) of the plot after flowering, and time to flowering

(FT) in days from January 1st. For assessment of the prediction bias protein content (PC) and oil content (OC) were considered. Both were adjusted for standard water content of 9%. A total of 188 plots in LAU 2020 were excluded due to water logging. Using the model

$$y_{ijklmn} = \mu + g_i + e_j + r_{k(ej)} + c_{l(ej)} + a_m + l_n + (g_i \times e_j) + \varepsilon, (1)$$

the adjusted entry means for each trait of the remaining data were calculated using the R package ‘lme4’ (Bates et al. 2015). This model was selected following Roscher-Ehrig et al. (2024) in order to replicate their results. Here, y_{ijklmn} refers to the observed phenotypic data of each block, μ to the overall mean, and g_i to the genotypic effect. Both μ and g_i were defined as fixed effects and formed together the adjusted entry mean. The environmental effect e_j , the effect of the k^{th} row ($r_{k(ej)}$), and the l^{th} column ($c_{l(ej)}$) within the j^{th} environment were treated as random effects. Additionally, the random effects a_m for the m^{th} year, l_n for the n^{th} location, and the genotype-by-environment interaction effect ($(g_i \times e_j)$) between the i^{th} genotype and j^{th} environment were included as random effects. ε represents the random residual error. The calculations were conducted by using the ‘lme4’ package in R.

NIRS data

NIR reflectance was measured using the SpectraStar 2500X RTW (KPM Analytics, Westborough, MA, USA) from harvested seeds of the hybrids grown at 5 locations and from 251 pollinators grown at RHH, using bulked samples. The wavelengths ranged from 680 nm to 2500 nm, with a 1 nm interval. Each measurement was taken twice and averaged. Subsequently, the reflectance values were adjusted using the same mixed linear model as for the other phenotypic data (1). Subsequently, the SG smoothing algorithm was applied using a cubic polynomial function, a window size of 37, and the first derivative. These settings were chosen according to Roscher-Ehrig et al. (2024) to maintain comparability of the results between both studies. This preprocessing step was implemented using the R package ‘prospectr’ (Stevens and Ramirez-Lopez 2025). The resulting profile was then centered and scaled.

NIRS-BLUP and genomic BLUP model

The underlying models for both PS and GS had the form

$$y = X\beta + Zu + e, (2)$$

where y is the vector of adjusted entry means, and X is the design matrix for the fixed effects β . The fixed effects included the fixed intercept and an effect for the two male sterile mothers of the hybrids. Z refers to the design matrix of the random effects, and u is the vector of the random genotype effects. e represents the random residual term. The

random effects are assumed to follow a normal distribution with $u_{NIRS} \sim N(0, G_{NIRS}\sigma_{NIRS}^2)$, for the NIRS-BLUP, and $u_{SNP} \sim N(0, G_{SNP}\sigma_{SNP}^2)$ for the genomic BLUP (GBLUP) model. The random residual is assumed to be normal distributed so that $e \sim N(0, I\sigma_e^2)$.

The phenomic similarity matrix was calculated according to Rincent et al. (2018) as follows:

$$G_{NIRS} = \frac{WW'}{q}, (3)$$

where W contains the processed NIRS data and has the dimensions $n \times q$, where n is the number of genotypes and q the number of wavelengths.

The genomic relationship matrix was calculated according to VanRaden (2018) as follows:

$$G_{SNP} = \frac{MM'}{2\sum p_i(1-p_i)}, (4)$$

where the elements of M are $(0-2p_i)$ for homozygous allele A, $(2-2p_i)$ for homozygous allele B and $(1-2p_i)$ for the heterozygous. The dimensions of M are $n \times m$, with n genotypes and m markers.

In random cross-validation, 80% of the data was randomly selected to estimate the genotypic effects using the NIRS-BLUP and GBLUP model, respectively. These estimated effects were then applied to the remaining 20% of the genotypes, which served as the prediction set. To evaluate the model's performance, the Pearson correlation coefficient (r) was calculated to quantify the prediction accuracy. The correlation coefficient was computed as:

$$r = \frac{cov(\hat{u}, \tilde{u})}{\sigma_{\hat{u}} \sigma_{\tilde{u}}}, (6)$$

where $cov(\hat{u}, \tilde{u})$ denotes the covariance between estimated and predicted genotypic value of the training and prediction sets, and $\sigma_{\hat{u}}$ and $\sigma_{\tilde{u}}$ represent their respective standard deviations (Pearson and Lee 1903; González-Recio et al. 2014). This cross-validation procedure was repeated 200 times, and the median of the resulting r values was used as an indicator of the model's prediction accuracy. All computations were performed using the R package 'sommer'. For within-generation prediction, both the NIRS data and phenotypic data from seed harvested from the hybrid genotypes were used. In contrast, across-generation prediction was performed by combining NIRS data from the father lines with phenotypic data from the hybrids. The paternal NIRS data were available for 251 genotypes grown at the RHH location. However, for 9 of the pollinators, no NIRS data was available. Consequently, all hybrids lacking a corresponding pollinator with available NIRS data were excluded from the analysis for predictions based on paternal NIRS data.

Heritability filtering

To determine the optimum wavelength heritability required for the highest prediction accuracy, the broad-sense heritability was calculated. Broad-sense heritability was calculated according to the formula of Piepho and Möhring (2007) for unbalanced data:

$$\bar{H}_2 = \frac{\sigma_g^2}{\sigma_g^2 + \bar{v}_{BLUE}/2}, (7)$$

where σ_g^2 refers to the genotypic variance and $\bar{v}_{BLUE}$ refers to the mean variance of the difference of two treatment means. Heritability was calculated from the raw, unadjusted,

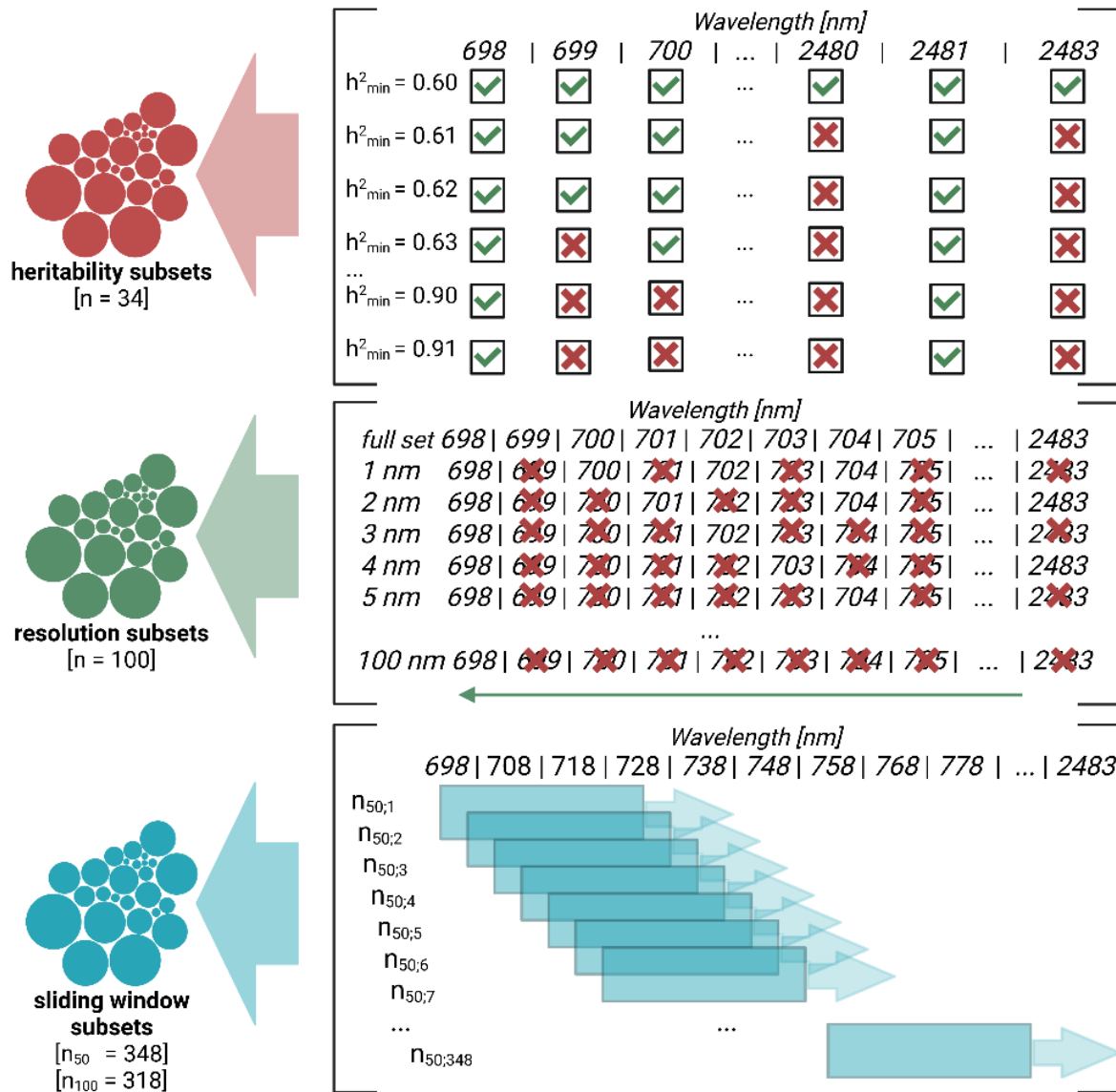


Fig. 1 Schematic overview of the filtering procedure. For heritability filtering, wavelengths were chosen above a given wavelength heritability threshold. Resolution filtering was done by increasing the intervals between each wavelength, starting always with the first wavelength. The sliding window approach yielded 238 and 318 subsets for a window size of 50 nm and 200 nm respectively. The window was moving towards maximum wavelength of 2483 nm after SG smoothing with a step size of five wavelengths for each. subset. The graph was made with Biorender ([citation URL](#)).

and unsmoothed NIRS data to account for the full variance. σ_g^2 was obtained from a full random model similar than that was used to adjust for the trial effects as described above (1) with exception that g_i was taken as random effect in this case. This analysis was conducted using the R package ‘lme4’. To compute $\bar{v}_{BLUE}$, the same model structure was applied; however, both the overall mean (μ) and the genotype effect (g_i) were treated as fixed effects.

Subsets of the adjusted and smoothed NIRS data were subsequently generated based on heritability thresholds. A continuous sequence of thresholds, ranging from the minimum to the maximum observed heritability in increments of 0.01, was employed to partition the data. Heritability estimates were derived from the raw hybrid NIRS data. These estimates were then used to generate subsets from the adjusted NIRS data for within-generation prediction, and from the adjusted pollinator data for across-generation prediction (Fig. 1). For each resulting subset of filtered NIRS data, model 2 was applied with subsequent assessment of prediction accuracy as described above.

Sliding window filtering

In this approach, sliding windows were applied across the entire spectral range, using window sizes of 50 and 200 nm, respectively. The windows were shifted incrementally from shorter to longer wavelengths with a step size of 5 nm (Fig. 1). For each resulting subset, model 2 was applied with subsequent assessment of prediction accuracy as described above.

Resolution filtering

This approach was designed to increase the spacing between selected wavelengths. Specifically, every i^{th} wavelength was selected, beginning consistently with the first wavelength in the spectrum. The interval i was progressively increased by 1 nm for each subset, up to a maximum spacing of 100 nm (Fig. 1). As the selection always commenced from the first wavelength, the specific wavelengths included in the subsets differed, even when the total number of wavelengths per subset remained constant in many cases. Also, here the prediction accuracy was assessed with a random cross validation as described above.

Random Wavelengths

For each subset from the previously described approaches that yielded the highest prediction accuracy, 200 randomly selected wavelength sets were generated, each containing the same number of wavelengths as the corresponding optimal subset. Each of these random sets was evaluated using 200-fold cross-validation, following the procedure outlined above. An overview of the pre-processing steps and the prediction approaches is presented in Figure 2.

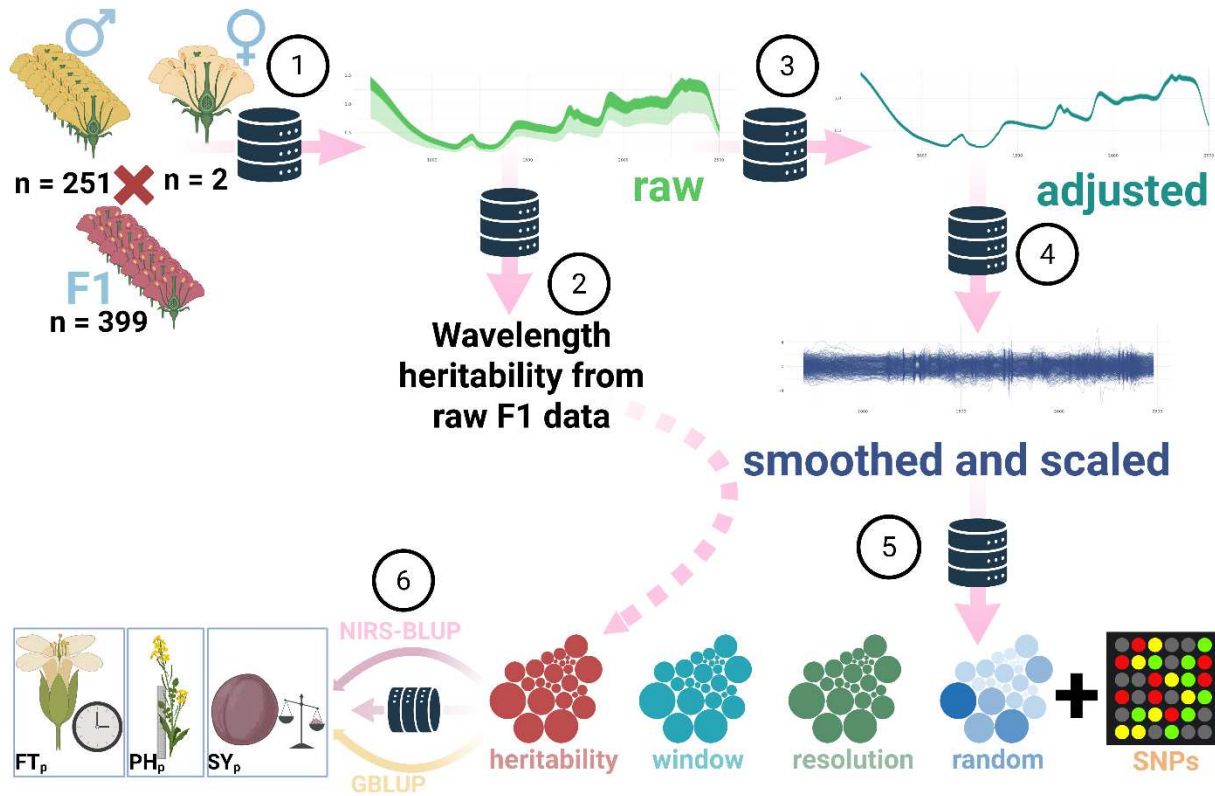


Fig. 2 Overview of the prediction pipeline: A describes the pre-processing of the NIR data, B describes the prediction approaches. (1) NIRS acquisition from the father and hybrid population. (2) Heritability estimation according to Piepho and Möhring (2007) using raw wavelengths. (3) Adjustment of the wavelengths. (4) SG-smoothing, centering and scaling of the adjusted wavelengths. (5) Formation of the different subsets according of the previously estimates heritability, the sliding window, the step-wise decrease of resolution and random selection. (6) Prediction and 200-fold cross validation for each subset and GBLUP. Diagram drawn with Biorender ([citation URL](#)).

Prediction bias and collinearity assessment

Collinearity among wavelengths was assessed by calculating r between all pairs of possible wavelength combinations. Prediction bias was assessed by calculating r of all possible pairwise combinations between the adjusted entry means of the traits (SY, PH, FT, PC, OC), as well as the predicted values of SY (SY_p), PH (PH_p) and FT (FT_p). PC and OC were included to determine whether these compounds differ in their correlation to predicted traits and the adjusted entry means of the corresponding traits. The term *bias* refers to the difference in r between the adjusted entry mean or its predicted counterpart and PC or OC, respectively.

Results

The SG smoothing procedure resulted in a total of 1786 wavelengths, ranging from 698 nm to 2483 nm. These processed spectral markers were subsequently subjected to the

analytical procedures outlined above (Fig. 2). The resulting spectral data were then used to construct G_{NIRS} employed in the NIRS-BLUP model.

Prediction with heritability thresholds for the wavelength

In the first approach, predictions were performed using only those wavelengths whose heritability exceeded a predefined threshold. For within-generation predictions, the wavelengths were derived from hybrid seeds across all environments. The adjusted phenotypic values for SY, PH, FT and the preprocessed NIR reflectance values were used to train the model. To investigate the influence of wavelength heritability on across-generation prediction, wavelengths obtained from the paternal genotypes cultivated in RHH were also employed as predictors.

The heritability of each individual wavelength was calculated as previously described. The mean heritability of the complete set of wavelengths derived from the hybrid genotypes was 0.88, with values ranging from 0.60 to 0.94. Notably, 1,711 wavelengths (94%) exhibited a heritability greater than 0.80. The previously described algorithm, which iteratively eliminated wavelengths based on their heritability, resulted in 34 distinct subsets for both father and hybrid NIRS data.

Figure 3 summarises the prediction accuracies of all investigated traits, using different subsets of wavelengths depending on their respective wavelength heritabilities. For the trait SY, the best subset for within generation prediction had a heritability threshold of $h^2=0.87$ and a median prediction accuracy of 0.55 (Fig 3, A;D). In across generation predictions the threshold was 0.88 and the median prediction accuracy at 0.33 (Fig. S1 A;D). The accuracy of the full sets also decreased from 0.54 to 0.32, whereas the random set accuracy decreased from 0.54 to 0.36 from within- to across-generation prediction (Fig. 3, A;D ; Fig. S1 A;D). The median prediction accuracy with the GBLUP model was 0.38 in within-generation prediction when including all genotypes, and 0.42 without the missing father genotypes (Fig. 3; Fig. S1).

Table 1 shows the used threshold for the best subsets and the number of remaining wavelengths. For the traits SY and PH the number of wavelengths is lower in across generational prediction, prediction of FT seems to require less wavelengths when predicted within a generation. However, the difference of used wavelengths for the best subset between across and within a generation is lower when predicting FT than predicting SY and PH.

For both SY and PH, the PS exceeded the GS accuracy when predicting within a generation. The GBLUP model had an accuracy of 0.38 and 0.25 for SY and PH, respectively. In contrast, the NIRS-BLUP models had an accuracy of 0.54, for all cases, except for the prediction of SY with the best subset, that achieved an accuracy of 0.55 (Fig. 3 A;B;D;E). For FT, the accuracy of the GBLUP model was 0.65 when predicted within

a generation. Here, the model had higher prediction accuracy than the corresponding NIRS-BLUP model that had an accuracy of 0.59 (Fig. 3 C;F).

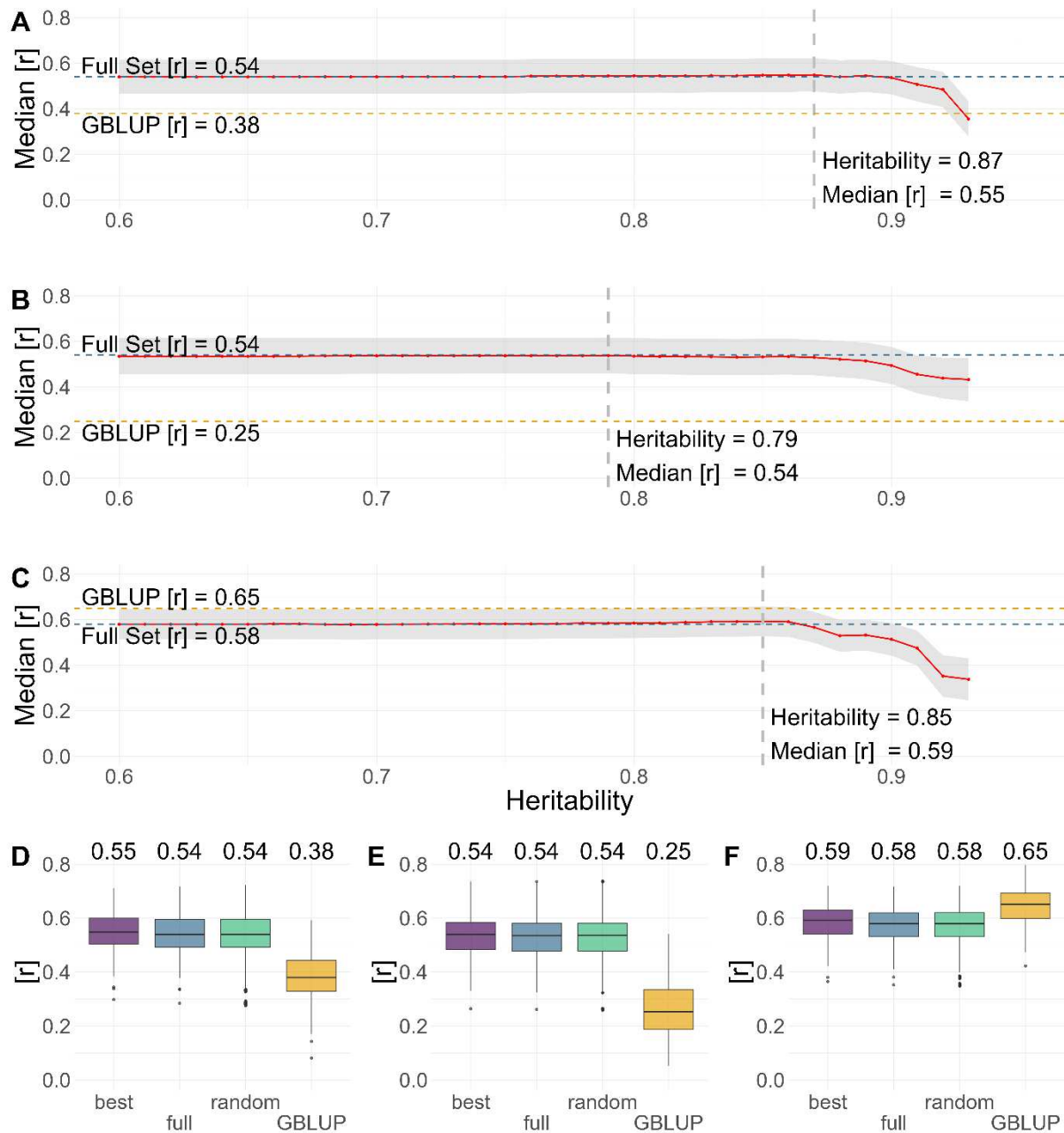


Fig. 3 Within-generation prediction accuracy [r] of (A; D) seed yield, (B;E) plant height and (C;F) flowering time after heritability filtering. The grey dashed line indicates the best subset along with its prediction accuracy and the applied heritability threshold. The blue dashed line represents the prediction accuracy achieved using the full spectral dataset, while the orange dashed line shows the prediction accuracy of the corresponding GBLUP model. D, E and F show the median prediction accuracy of the best subset, the full set, the random subset and the GBLUP model with its standard deviation, respectively.

Table 1 Overview of all wavelength reduction approaches and corresponding trait heritability (*heritability estimates were obtained from Roscher-Ehrig et al. (2024).).

	SY		PH		FT	
	<i>within</i>	<i>across</i>	<i>within</i>	<i>across</i>	<i>within</i>	<i>across</i>
<i>$\bar{H}^2$-wise</i>						
Best subset threshold	0.87	0.89	0.79	0.90	0.85	0.82
Median r	0.55	0.33	0.54	0.20	0.59	0.55
No. of wavelengths	1376	975	1716	755	1570	1661
<i>Resolution-wise</i>						
Best subset interval in nm	92	81	58	74	88	35
Median r	0.59	0.35	0.56	0.21	0.61	0.56
No. of wavelengths	20	23	31	25	21	51
<i>Sliding window</i>						
Mean wavelength	1518	1003	993	1073	1088	1793
Median r	0.60	0.34	0.54	0.20	0.55	0.42
No. of wavelengths	200	50	200	50	200	200
Full set r	0.54	0.32	0.54	0.20	0.58	0.54
GBLUP r	0.38	0.42	0.25	0.26	0.65	0.66
Trait heritability*	0.49	-	0.47	-	0.63	-

The predictions across generation showed remarkable lower prediction accuracy with a NIRS-BLUP model when targeting the traits SY and PH (Fig. S1 A;B;D;E). However, the across generational prediction of the trait FT reached an accuracy of 0.54 and 0.55 for the full set and the best subset, respectively. Only the random choice was relatively low with an accuracy of 0.37 (Fig. S1 C;F).

For all predictions, within- and across-generations, the change of accuracy between the single subsets with increasing wavelength heritability thresholds was minimal until a threshold, close to the heritability maximum of 0.94 was reached. In most cases, a remarkable decrease of prediction accuracy occurred close after the threshold increase further (Fig. 3 , Fig. S1).

Prediction with resolution filtering

In this approach, the intervals between the wavelengths increased by 1 nm for each subset. Figure 4 summarises the prediction accuracies of all investigated traits for within-generation predictions, using different subsets of wavelengths depending on their respective spectral resolution. In within-generational prediction, the best subset for SY with an interval of 92 nm exceed the full set prediction and resulted in an accuracy of 0.59 (Fig. 4 (A,D)).

In across generation prediction, the best subset for SY with intervals of 81 nm reached an accuracy of 0.35. The corresponding GBLUP yielded an accuracy of 0.42 (Fig. S2 (A,D)). Comparable results were observed for the traits PH and FT. For PH, accuracies of 0.56

(within-generation) and 0.21 (across-generation) were obtained. In both cases, the random choice did not surpass the best subset. For PH, the GBLUP model was 0.31 lower in within-generation prediction but comparable in across-generation prediction (Fig. 4 (B,C,E,F), Fig. S2(B,C,E,F)).

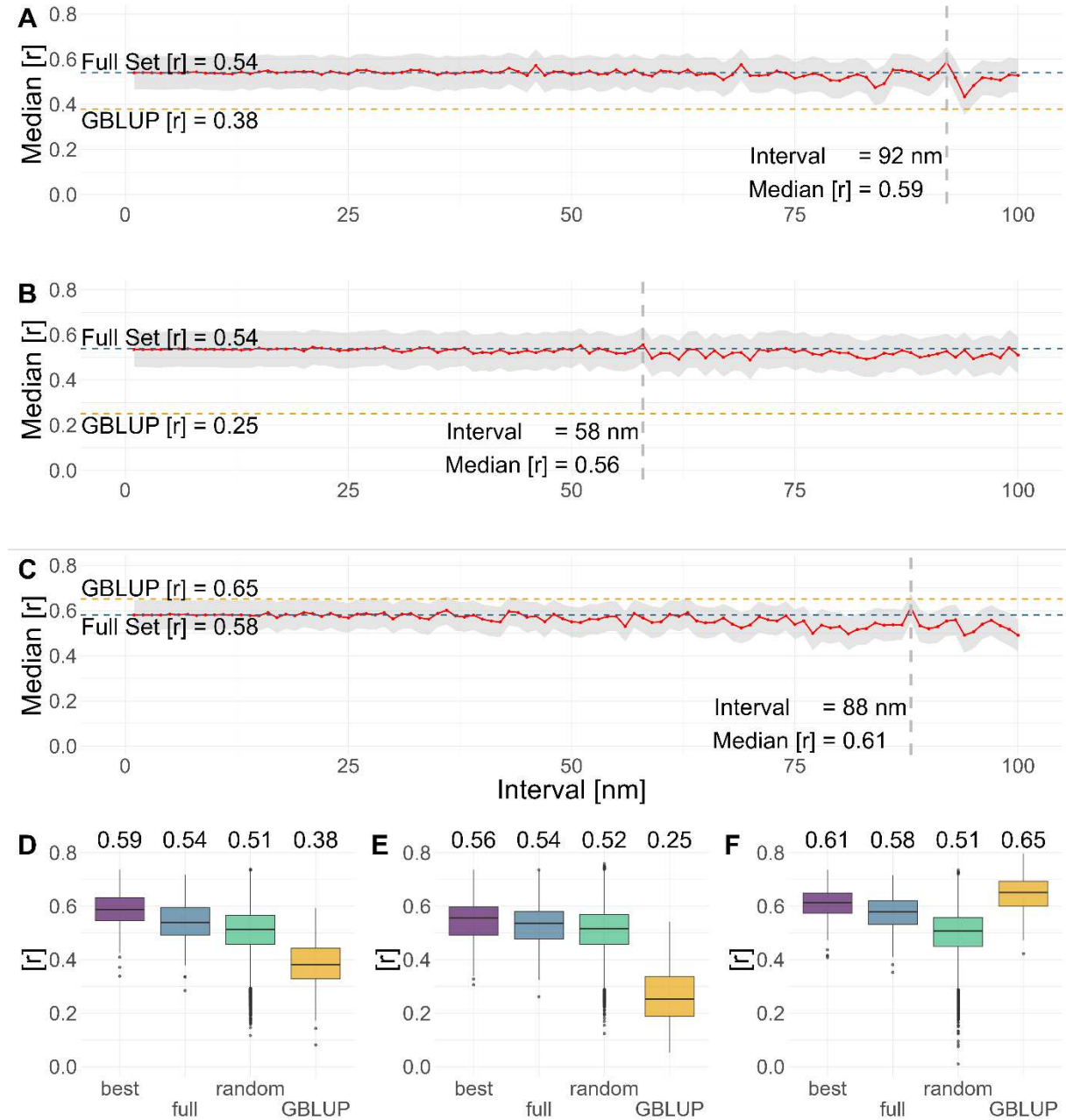


Fig.4 Within-generation prediction accuracy of (A; D) seed yield, (B;E) plant height and (C;F) flowering time after resolution filtering. The grey dashed line indicates the best subset along with its accuracy and the interval between the wavelengths. The blue dashed line represents the accuracy achieved using the full spectral dataset, while the orange dashed line shows the prediction accuracy of the corresponding GBLUP model. D, E and F show the median prediction accuracy of the best subset, the full set, the random subset and the GBLUP model with its standard deviation, respectively.

For FT, the GBLUP model achieved a similar prediction accuracy compared to NIRS-based predictions within the generation (Fig. 4). Across the generation, the accuracy was 0.1 higher when using the SNP markers (Fig. S2). NIRS-based prediction surpassed the performance of the full spectral set when employing wavelength intervals of 88 nm and 35 nm for within- and across-generation predictions, respectively; however, the improvement did not exceed 0.03 in either case. In contrast, the use of randomly selected wavelengths led to a substantial decline in predictive accuracy, with median losses of 0.1 and 0.2 for within- and across-generation predictions, respectively (Fig. 4 ; Fig. S2).

The number of used wavelengths for the best subsets are shown in Table 1. In within generation prediction between 20 (1.1%) and 31 (1.7%) wavelengths out of 1786 resulted in the best prediction accuracy. When predicting across generations it's between 23 (1.3%) and 51 (2.9%) out of 1786.

Prediction with a sliding window

The sliding window was applied across the entire spectrum with window sizes of 50 nm (Fig.5-6) and 200 nm (Fig. S3-S4). For a window size of 50 nm, the best-performing windows did not exceed the accuracy of the full set in within-generation prediction (Fig. 5).

However, across generations, the 50 nm window outperformed the full set for SY (Fig.6). Notably, there are regions where the 50 nm window achieved higher prediction accuracies, which appear to be trait dependent (Fig. 5 (A–C)). With a 200 nm window, these regions largely disappeared. Particularly in across-generation predictions, window-specific peaks of prediction accuracy were reduced and nearly vanished when the window size was increased (Fig. S3 (A–C), Fig. S4 (A–C)). Only for FT were distinct peaks still relatively prominent in across-generation prediction with a 50 nm window (Fig. S4 C).

A 200 nm window exceeded the full set accuracy by 0.06 for SY in within-generation prediction (Fig. S3 (A,D)), but the opposite was observed across generations (Fig. S4 (A,D)).

The outperformance of the 200 nm window in within-generation prediction for SY was not evident for the other traits. Random choice performed similarly or slightly better than the full set in within-generation prediction (Fig. 5; Fig. S3). Across generations, however, the full set yielded slightly higher prediction accuracies, except for SY (Fig. 6, Fig. S4).

Collinearity and bias

To determine degree of collinearity between the wavelengths the pairwise r between each wavelength was calculated. Figure 7 shows heatmaps of the NIR reflection values before and after SG smoothing. Prior to smoothing, pairwise correlations were predominantly positive, ranging from 0 to 1. Two distinct clusters were visible in the top right and bottom left of the heatmaps for the adjusted hybrid and father NIRS. After SG smoothing, the

correlation patterns changed markedly, spanning the full range from -1 to 1 . The previously observed clusters disappeared (Fig. 7), and the large cluster of correlated wavelengths was replaced by multiple smaller clusters.

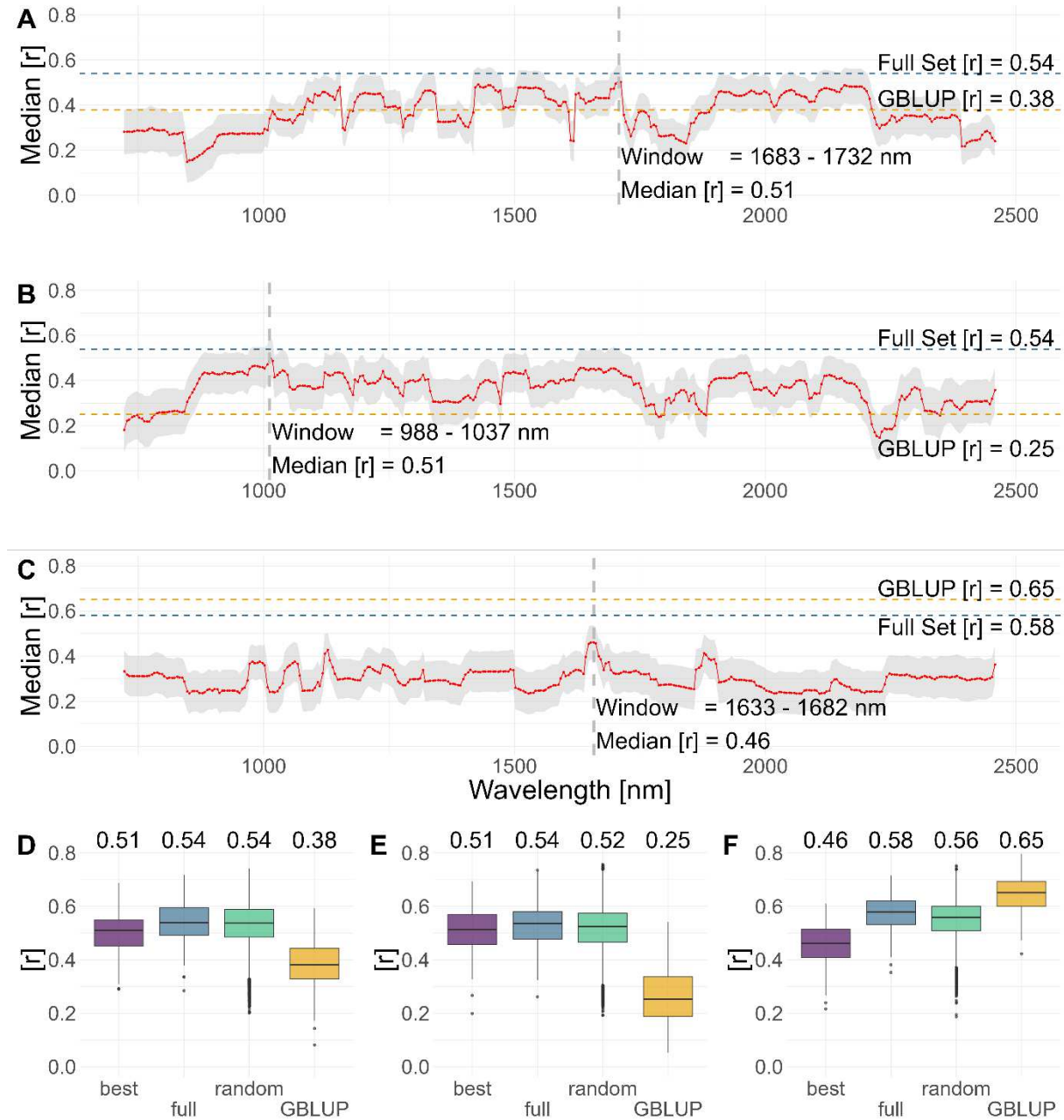


Fig. 5 Median prediction accuracy of within-generational predictions. Each red dot represents the median prediction accuracy of a single 50 nm window at the mean wavelength of the window for the traits seed yield (A), plant height (B), and flowering time (C). The grey dashed line indicates the best subset, while the full set accuracy and GBLUP accuracy are indicated by the blue and yellow lines (A–C), respectively. Boxplots on the right-hand side (D–F) show the corresponding best, full, random, and GBLUP sets. Bars represent the standard deviation, and lines indicate the median prediction accuracy (values given above) obtained after 200-fold cross validation.

Moreover, wavelengths that were previously mostly positively correlated now exhibited both negative and positive correlations (Fig. 8). The transition from negative to positive

correlation appeared abrupt, resulting in sharp colour changes in the heatmaps (Fig. 7). The bottom-left cluster was less affected by SG smoothing. Overall, SG smoothing increased the complexity of the correlation structure among wavelengths and reduced multicollinearity. Figure 8 illustrates that, following SG smoothing, the distribution of pairwise correlations became much broader, whereas before smoothing it was strongly peaked around 0.9.

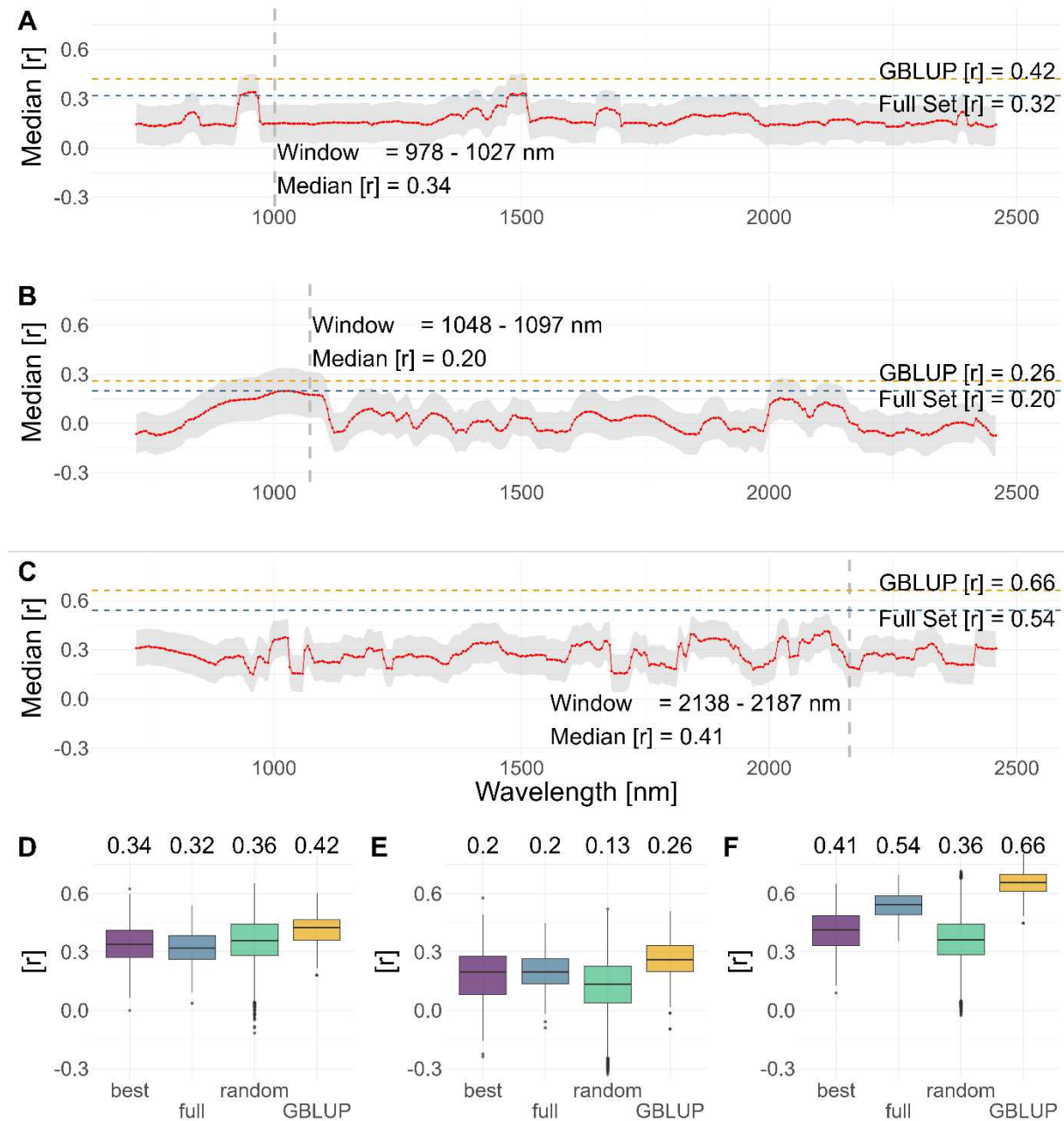


Fig.6 Median prediction accuracy of across-generational predictions. Each red dot represents the median prediction accuracy of a single 50 nm window at the mean wavelength of the window for the traits seed yield (A), plant height (B), and flowering time (C). The grey dashed line indicates the best subset, while the full set accuracy and GBLUP accuracy are indicated by the blue and yellow lines (A–C), respectively. Boxplots on the right-hand side (D–F) show the corresponding best, full, random, and GBLUP sets. Bars represent the standard deviation, and lines indicate the median prediction accuracy (values given above) obtained after 200 fold cross validation.

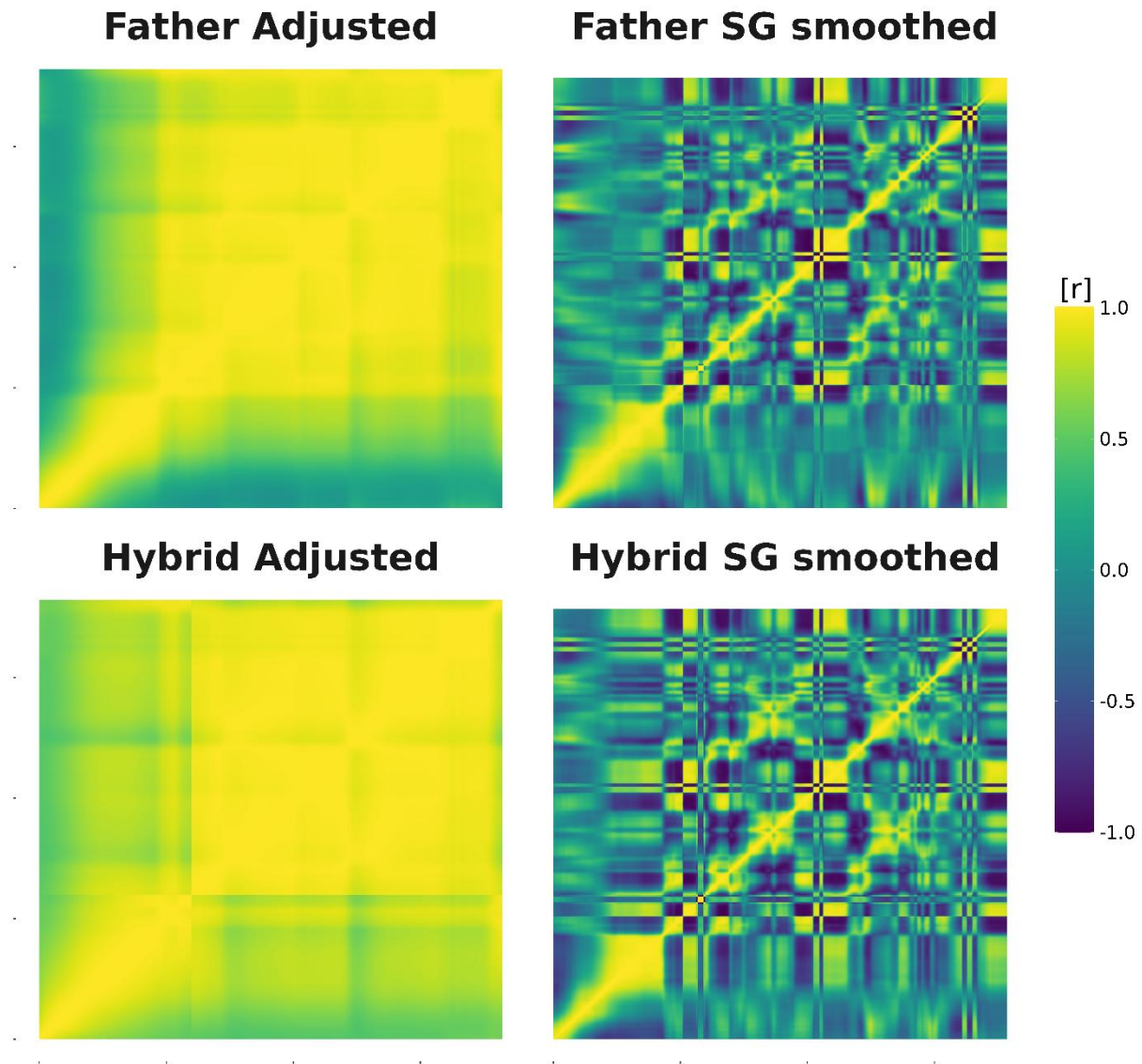


Fig. 7 Heatmaps of the pairwise correlations between all possible combinations of two adjusted entry means of the wavelengths before and after Savitsky-Golay smoothing. Blue indicates negative correlation, yellow indicates positive correlation.

Figure 9 shows that the correlation between SY and PC was -0.33 , representing the realized correlation between these traits, whereas the correlation between SYp and PC was -0.57 , representing the predicted correlation and indicating an introduced bias. A similar pattern was observed for SY and OC (-0.11) and for SYp and OC (-0.19). Comparable trends were found for PH and PC (-0.28) and PHp and PC (-0.57). In contrast, FT and FTp exhibited only very weak correlations with both PC and OC. The same holds true for the correlations of PH and PHp with OC.

Discussion

This study aimed to assess the quality requirements of individual wavelengths used in PS. Therefore, the heritability threshold for selected wavelengths was gradually increased to optimize prediction accuracy. In parallel, spectral resolution was systematically decreased by widening the intervals between individual wavelengths. Lastly, relevant

spectral regions were identified using a sliding window approach, which examined discrete sections of the spectrum in turn.

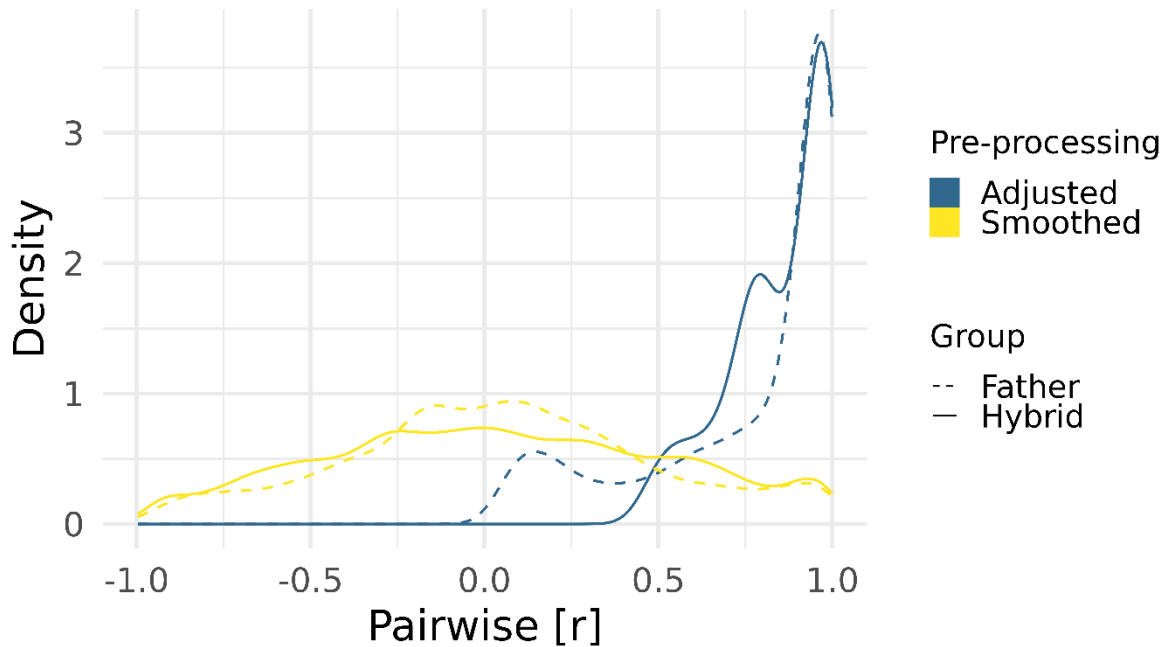


Fig. 8 Density distribution of the pairwise correlation of the adjusted entry means of the wavelengths and the additionally smoothed values, obtained from the pollinator seeds and the hybrid seeds.

Prediction was either based on NIRS spectra obtained from seeds harvested from hybrids (within generation prediction) or from seeds harvested from parental lines (across generation prediction). When prediction is performed within a generation, PS outperforms GS in these rapeseed hybrid trials, as previously investigated by Roscher-Ehrig et al. (2024). However, the optimal strategy for maximizing the prediction accuracy of PS models remains under debate. For example, Braun et al. (2025) optimized the SG smoothing, as initially suggested by Robert et al. (2022a). Other authors used multiple tissues such as wood, leaves (Rincen et al. 2018; Brault et al. 2022) and grains (Lane et al. 2020) for predictions. Further, different model approaches were tested with little gain in accuracies in most cases. Penalized regression approaches that incorporate feature selection, namely Least Absolute Shrinkage and Selection Operator (LASSO), have been shown not to outperform NIRS-BLUP in predicting target traits (Zhu et al. 2021; Roscher-Ehrig et al. 2024; Adunola et al. 2024). However, Lopez-Cruz et al. (2020) showed that applying dimensionality-reduction or feature-selection methods can lead to notable gains in predictive accuracy, when applied on hyperspectral imaging data. Moreover, Robert et al. (2022a) propose, to use wavelengths with high heritability, since these are more representative of genetic variances. Even though some studies have also investigated specific wavelengths of importance, including Braun et al. (2025) and Robert et al. (2022a), none of these studies addressed whether it is necessary to include all wavelengths in the model a priori.

FTp	-0.05	0.2	0.68	0.04	-0.04	-0.13	0.27	1
PHp	0.25	0.61	0.17	-0.03	-0.5	0.41	1	0.27
SYp	0.62	0.24	-0.07	-0.19	-0.57	1	0.41	-0.13
PC	-0.33	-0.28	-0.03	-0.53	1	-0.57	-0.5	-0.04
OC	-0.11	-0.02	0.05	1	-0.53	-0.19	-0.03	0.04
FT	-0.02	0.24	1	0.05	-0.03	-0.07	0.17	0.68
PH	0.38	1	0.24	-0.02	-0.28	0.24	0.61	0.2
SY	1	0.38	-0.02	-0.11	-0.33	0.62	0.25	-0.05
	SY	PH	FT	OC	PC	SYp	PHp	FTp

Fig. 9 Pairwise correlations between the realized traits seed yield (SY), plant height (PH), flowering time (FT), oil content (OC), protein content (PC) and the predicted counterparts SYp, PHp and FTp.

Wavelength heritability is an indicator for environmental influence on the wavelength

Montesinos-López et al. (2017) and Robert et al. (2022a) recommended the use of wavelengths with heritability values above 0.5, as these are expected to contribute meaningful information to the prediction model. To validate the relevance of wavelength heritability, in this study wavelengths were progressively eliminated based on their heritability. This approach reduced the number of wavelengths without compromising the median prediction accuracy, with reductions of required spectra ranging from 7% for FT to 58% for PH, predicted across generations. Within generations, the reductions ranged from 4% up to 23% (Table 1). However, it is important to note that all wavelengths in this study exhibited heritability values ≥ 0.6 , with a mean heritability of 0.88. Consequently,

the degree of wavelength reduction was likely smaller than what might be observed in datasets with a wider distribution of wavelength heritability. Furthermore, the distribution of heritability across wavelengths appears to depend on the tissue from which reflectance was measured. Wavelengths derived from wood and leaf samples generally exhibit lower heritability compared to those from grains (Rincent et al. 2018; Brault et al. 2022). This difference can be explained by the strong NIR absorbance of water and the resulting dilution effects on other compounds (Büning-Pfaue 2003; Clevers et al. 2008). Because tissue water content is strongly influenced by environmental factors such as temperature, soil moisture, and time of day (Allen et al. 1998), wavelengths measured from fresh plant material are likely to display greater environmental and genotype x environment variance (Bowman 1989; Li et al. 2001), and thus lower broad-sense heritability.

Nonetheless, the findings here suggest that a heritability threshold of approximately $h^2=0.8$ or higher may be more appropriate for capturing all relevant wavelengths in seeds for prediction in this rapeseed trial, as lowering the threshold did not result in increased prediction accuracy. However, none of the wavelength subsets yielded an accuracy gain greater than 0.01 compared to predictions based on the full spectrum. This pattern was consistent across both within-generation and across-generation predictions (Fig. 3, Fig. S1).

When using the same number of wavelengths but selecting them randomly across the entire spectrum, only minor differences in prediction accuracy are observed within a generation. However, in across-generation prediction, accuracy increases by 0.03 for SY but decreases by 0.06 and 0.18 for PH and FT, when comparing the best-performing subset to the random subset (Fig. S1). This decline in accuracy for FT and PH is likely attributable to wavelength heritability, as wavelengths with higher heritability capture a greater proportion of genotypic variance, which can be transmitted to the next generation. (Piepho and Möhring 2007; Schmidt et al. 2019). Consequently, only wavelengths with high heritability contribute meaningfully to predicting offspring performance based on the paternal phenotype. In contrast, randomly selected wavelengths are less likely to represent genotypic variance. Therefore, wavelength heritability is a critical factor in determining whether a wavelength contributes meaningful information to the prediction model.

This pattern does not apply to the trait SY, for which randomly selected wavelengths yielded higher prediction accuracy across generations. This may be explained by the fact that SY is predominantly influenced by the plant, which supplies seed components. A strong maternal effect is well-documented for related traits such as OC, PC, and thousand-kernel weight (McVetty and Pinnisch 1994; Wang et al. 2010; Li et al. 2015). Hence the maternal effect is included in the model for hybrid prediction. However, also the fathers' seed components are strongly influenced by the plant where they were growing on. Since NIR reflectance was measured on the fathers' seeds, the effect of the

paternal plant on the seed components creates bias towards it. As such, it is likely that the father's NIRS data lacks key maternal information necessary for accurately predicting hybrid SY, even when the selected wavelengths exhibit high broad-sense heritability.

Sliding windows uncover spectral regions of importance

To find spectral regions which significantly affect the prediction accuracy, predictions were conducted using a sliding window approach with varying window sizes. The results clearly indicate the presence of multiple important regions across the spectrum. A smaller window size of 50 nm revealed several regions of importance, distributed across the entire spectrum in within-generation predictions. Notably, a 200 nm window yielded the highest prediction accuracy across all traits and produced a smoother accuracy curve across the sliding windows. This suggests that predictive information is not confined to narrow bands but is instead distributed across broader spectral ranges.

In across-generation predictions for the trait SY, two distinct spectral regions produced substantially higher prediction accuracies than others (Fig. 5). This pattern contrasts sharply with the within-generation predictions using the same window size, where a broader set of influential regions becomes apparent. However, more similar patterns were observed for PH and FT. Interestingly, while randomly selected subsets outperformed the best spectral regions for SY in across-generation predictions, the opposite was true for PH and FT (Fig. 5, Fig. S4). This discrepancy may be attributed to the pronounced maternal effect on SY, as discussed previously. In this case, maternal effects diminish the predictive power of targeted wavelength selection, making random subsets more effective at capturing residual information relevant to offspring SY performance.

The comparatively good prediction accuracy for FT across all approaches may be attributed to the generally high heritability of this trait (Table 1) (Roscher-Ehrig et al. 2024). Consequently, particularly in across-generation predictions, a substantial amount of information is likely transmitted from the father to the offspring.

Prediction with decreased resolution can enhance model accuracy

When reducing the resolution by increasing the intervals between the selected wavelengths, an interesting pattern emerges: the best subsets use only a very small number of wavelengths. In the case of within-generation prediction for SY, only 20 wavelengths—approximately 1.1% of the total—are utilized (Table 1). Additionally, prediction accuracy appears to oscillate around the prediction accuracy of the full set and decreases only gradually, on average, as the intervals widen (Fig. 4). Notably, the best subsets emerge only at relatively broad intervals between 58 nm (PH) and 92 nm (SY), and the oscillating pattern seems to become more pronounced as the intervals increase (Fig. 4). In across-generation prediction, similar patterns were observed. The best subsets

occurred at window sizes ranging from 35 nm for FT to 81 nm for SY. In most studies, NIRS data with a wavelength spacing of 1 nm (Zhu et al. 2022; Brault et al. 2022; Roscher-Ehrig et al. 2024) or 2 nm (Rincen et al. 2018; Lane et al. 2020; Robert et al. 2022a; Braun et al. 2025) were used. Here, it could be shown that even lower spectral resolutions are suitable for PS, at least when the heritability of the wavelengths is relatively high.

Similar results were observed when using hyperspectral images obtained from drone-mounted sensors. L1-penalized selection indices (L1-PSI) achieved the best outcomes through wavelength selection that was independent of the plant's developmental stage (Lopez-Cruz et al. 2020). In that study, only a few wavelengths remained in the model, meanwhile increasing prediction accuracy. This parallels the findings of the present study: not only was the number of wavelengths substantially reduced, but prediction accuracy can increase relative to the full-set model. These results indicate that selecting informative markers can achieve equal or even superior predictive power compared with a NIRS-BLUP model that incorporates all wavelengths. However, as noted earlier, Roscher-Ehrig et al. (2024) demonstrated that LASSO penalization and feature selection do not improve prediction accuracy in this dataset. Consequently, the optimal interval identified here suggests that accuracy gains arise not from sparse feature selection per se, but from capturing the most informative spectral regions while excluding redundant wavelengths that may otherwise introduce noise into the model.

Redundancy due to multicollinearity of wavelengths

Given the high pairwise correlations among wavelengths (Fig. 7), which indicate substantial collinearity, the information loss from randomly excluding highly collinear wavelengths is likely marginal (Hocking and Pendleton 1983). However, SG smoothing reduces the degree of multicollinearity (Fig. 8). Nevertheless, wavelengths are still likely to correlate with their neighbors, as indicated by the formation of distinct clusters in the heatmap of pairwise correlations. This finding underscores the high degree of redundancy among wavelengths and highlights the necessity of applying shrinkage to wavelength effects, as is commonly done in mixed linear models using the restricted maximum likelihood (REML) method. In REML, shrinkage toward the mean is governed by the ratio of error variance to marker variance (Henderson 1975; VanRaden 2008). However, this approach includes all wavelengths, which share common effects due to multicollinearity. While these wavelengths may not contribute additional information to the model, they can introduce bias, as their signal derives from the molecular composition of the seed. This composition, in turn, is influenced by numerous factors, including environmental effects (Pritchard et al. 2000; Werteker et al. 2010) and, as discussed above, maternal effects. Consequently, reflectance is a function not only of genotypic effects but also of environmental influences, genotype-by-environment interactions, and residual error. For this reason, wavelengths were pre-adjusted to account for trial and environmental effects prior to analysis. The reduced introduction of noise through the exclusion of redundant wavelengths likely explains the observed increase in prediction accuracy with wider

intervals, provided that all relevant spectral regions are represented by at least one representative wavelength.

Limitations of wavelength correlation relative to linkage disequilibrium

The concept of redundant markers is also well established in the context of genetic markers such as SNPs. Linkage disequilibrium (LD), defined as the deviation from the expected random inheritance of two alleles (Weir 1979), allows the correlation between two genetic markers to be considered a measure of LD (Zaykin et al. 2008). This principle has been applied to assign genetic markers to causal loci for specific traits in genome-wide association studies (GWAS) (Altshuler et al. 2008). Subsequently, haplotype-based predictions were shown to improve prediction accuracy, particularly when haplotypes were defined based on LD between markers (Won et al. 2020). These improvements appear to be trait dependent and differ between within- and across-generation predictions (Difabachew et al. 2023; Lin et al. 2024). Furthermore, the gain in accuracy is also dependent on both trait and model (Weber et al. 2023).

However, this direct relationship between genetic markers and causal loci does not apply to NIR markers. In the case of NIR, the link to the genotype is only indirectly observable via molecular proxies such as proteins, fatty acids, or carbohydrates. The strong correlations among closely spaced wavelengths likely result from molecular absorbance, as related molecules tend to absorb at nearby wavelengths (Chen et al. 2011). An indication that correlating proxies such as OC or PC contribute to this pattern is the structure of the subsets determined by the resolution filtering procedure: the first wavelength was always selected, followed by every i^{th} wavelength. As a result, neighbouring subsets contained the same number of wavelengths but were shifted along the spectrum. Examination of the best subsets and their neighbouring subsets shows that this shift affected prediction accuracy (Fig. 4). This suggests that specific wavelengths carry molecular proxy information, and their inclusion or exclusion, determined by the shift, impacts predictive performance, even though the total number of wavelengths remains constant. In this context, the correlation between wavelength and molecule can be viewed as analogous to LD, but at the level of seed compounds. However, this signal is obscured by inorganic molecules, particularly water, as discussed previously. Hence, forming wavelength blocks analogous to LD-based haplotypes could reduce environmental noise and thereby increase prediction accuracy, similar to the improvement observed in this study when selecting the 20 most important wavelengths based on the resolution filtering procedure, while neglecting the rest.

Molecular proxies can create bias depending on the trait

Recently, it was shown that NIRS-based predictions can create bias toward molecular proxies that correlate with the target trait (Dallinger et al. 2023). In this study, one such

proxy is PC, which negatively correlates with SY (Fig. 9). Because SY is associated with seed compounds of high abundance, such as proteins and oils (Degenhart and Kondra 1984; Variath et al. 2009; Wang et al. 2010; Chen et al. 2011), its prediction likely benefits from the absorbance properties of these correlated molecular proxies. Similarly, PH shows a negative correlation with PC (-0.28 ; Fig. 9), suggesting that PC may also act as a meaningful proxy for predicting this trait. Since NIR light is absorbed by chemical bonds in molecules such as proteins and fatty acids (Williams et al. 1985; Jiang et al. 2007), predictions are likely biased toward those compounds. This relationship is evident when comparing correlations between predicted trait values and their associated proxies. While the realized correlation between SY and PC was -0.33 , the correlation between SYp and PC was notably stronger at -0.57 . A similar pattern was observed for PH, where the correlation between PHp and PC reached -0.50 compared to the realized correlation of -0.28 . These findings highlight the role of correlated biochemical traits, such as protein content, in enhancing predictive power for agronomic traits like SY and PH. Consequently, NIR spectroscopy might also be more biased towards the specific field trials than genomic markers, because tissue composition, including molecular proxies, is strongly influenced by environmental conditions (Pritchard et al. 2000; Werteker et al. 2010) and is directly measured through NIR absorbance.

Interestingly, this consolidated correlation between predicted trait values and molecular proxies does not apply to FT. In this trial, the realized correlation between FT and PC was -0.04 , while the correlation between FTp and PC remained similarly low at -0.03 . For FT, numerous signalling molecules contribute to trait determination; however, their NIR-based signal may be obscured by noise due to their low concentrations (Chakraborty et al., 2022). This suggests that the concentration of molecules capable of introducing bias plays a critical role in NIR-based prediction.

Moreover, FT exhibits a generally high broad-sense heritability (0.63), which is greater than that of SY (0.49) and PH (0.47) (Roscher-Ehrig et al. 2024). This may explain its consistently higher prediction accuracy in both within- and across-generation scenarios (Fig. 3 and Fig. S1).

Roscher-Ehrig et al. (2024) compared predictions using an NIRS-BLUP and a SNP-based GBLUP on the same dataset. Their findings showed that, in within-generation prediction, the GBLUP model achieved higher accuracy for FT (0.65) than for SY (0.38) (Roscher-Ehrig et al. 2024). Since GBLUP do not benefit from proxy compounds, they yield lower accuracy for SY, whereas the NIRS-BLUP can leverage this source of information. As a result, GBLUP benefits from the highly genotypically determined nature of FT (Luan et al. 2009; Daetwyler et al. 2010). In contrast, the PS model for FT likely faces more noise and a lower abundance of relevant molecular proxies.

Although SY, FT, and PH are complex traits and regulated by multiple genetic and environmental factors (Schiessl et al. 2015; Li et al. 2016; Nowosad et al. 2016; Han et al.

2022), their dependency on seed compounds differs. These differences likely determine whether PS models benefit from strong molecular proxy effects meanwhile suffering from bias introduced through the environmentally influenced seed composition.

Wang et al. (2025) discussed the problem of introduced bias and argued that, for this reason, comparisons between GBLUP and NIRS-BLUP may not be meaningful. This may hold in cases where NIR markers introduce undesirable molecular bias. However, as long as predictions are not affected by such bias, as observed for FT, NIRS-BLUP models remain valuable for breeders, and comparisons of prediction accuracy can guide decisions on whether they should be implemented instead of GBLUP. Furthermore, this bias should be removed if NIRS values for genotypes are sampled like target traits across a representative sample of the target population of environments. Here the bias should be reduced by adjustment of the spectra for the experimental design and environments as applied in the current study. Nonetheless, it must be emphasized that bias in predicted values can be unfavourable for breeding, as NIRS-based selection may lead to the unintended negative selection of undesirable but correlated compounds. For example, when selecting for SY while keeping PC in rapeseed meal constant under a restricted index selection scenario, the response to selection would decrease significantly (Cunningham et al. 1970).

Environmental bias may also create challenges when using NIRS-based selection to develop cultivars for environments outside the target population of environments. Since genomic marker-based selection is more likely to capture pure additive effects, or at least better distinguish between $G \times E$ and genotypic effects, it is particularly useful for breeding broadly adapted cultivars. Conversely, environmental bias should not be a problem when the goal is to identify the best cultivar for the target population of environments. Hence, PS could be especially useful in a rapeseed hybrid breeding program in situations where limited seed availability leads to missing plots.

Summary

Given the previously discussed findings, the prediction accuracy of NIRS-BLUP models appears to depend on several factors. First, the quality of NIRS measurements strongly affects model performance. While wavelength heritability alone does not fully capture measurement quality, it serves as a useful indicator of genotypic influence, particularly when NIRS is performed on fresh tissue samples. However, strong collinearity among wavelengths leads to considerable redundancy, even when wavelength heritability is generally high. Second, the inheritance pattern of the target trait is critical. Traits with substantial maternal inheritance are particularly difficult to predict across generations, as paternal NIRS data fail to capture maternally derived seed compounds. Furthermore, molecular proxies such as PC and OC may introduce systematic bias into predicted values, especially when they are strongly correlated with the target trait. For these reasons, PS must be applied with caution, as it carries the risk of incorporating proxies

that may distort genetic evaluations and reduce selection response. Nonetheless, PS also offers notable advantages, including reduced costs and higher throughput due to inexpensive marker acquisition, as well as comparable or even superior prediction accuracies for certain traits. In particular, traits that are not strongly correlated with highly abundant molecular compounds, such as FT, represent promising candidates for PS, as they are less prone to bias and more likely to reflect true genetic potential.

References

- Adunola P, Tavares Flores E, Riva-Souza EM, et al (2024) A comparison of genomic and phenomic selection methods for yield prediction in *Coffea canephora*. The Plant Phenome Journal 7:e20109. <https://doi.org/10.1002/ppj2.20109>
- Allen LH, Valle RR, Jones JW, Jones PH (1998) Soybean Leaf Water Potential Responses to Carbon Dioxide and Drought. Agronomy Journal 90:375–383. <https://doi.org/10.2134/agronj1998.00021962009000030010x>
- Altshuler D, Daly MJ, Lander ES (2008) Genetic Mapping in Human Disease. Science 322:881–888. <https://doi.org/10.1126/science.1156409>
- Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using **lme4**. J Stat Soft 67:. <https://doi.org/10.18637/jss.v067.i01>
- Bowman WD (1989) The relationship between leaf water status, gas exchange, and spectral reflectance in cotton leaves. Remote Sensing of Environment 30:249–255. [https://doi.org/10.1016/0034-4257\(89\)90066-7](https://doi.org/10.1016/0034-4257(89)90066-7)
- Brault C, Lazerges J, Doligez A, et al (2022) Interest of phenomic prediction as an alternative to genomic prediction in grapevine. Plant Methods 18:108. <https://doi.org/10.1186/s13007-022-00940-9>
- Braun V, Zhu X, Meyenberg C, et al (2025) Phenomic Prediction Can Be Improved by Optimization of NIRS Preprocessing. Plant Breeding pbr.13265. <https://doi.org/10.1111/pbr.13265>
- Büning-Pfaue H (2003) Analysis of water in food by near infrared spectroscopy. Food Chemistry 82:107–115. [https://doi.org/10.1016/S0308-8146\(02\)00583-6](https://doi.org/10.1016/S0308-8146(02)00583-6)
- Chen GL, Zhang B, Wu JG, Shi CH (2011) Nondestructive assessment of amino acid composition in rapeseed meal based on intact seeds by near-infrared reflectance spectroscopy. Animal Feed Science and Technology 165:111–119. <https://doi.org/10.1016/j.anifeedsci.2011.02.004>
- Clevers JGPW, Kooistra L, Schaepman ME (2008) Using spectral information from the NIR water absorption features for the retrieval of canopy water content. International Journal of Applied Earth Observation and Geoinformation 10:388–397. <https://doi.org/10.1016/j.jag.2008.03.003>

- Cunningham EP, Moen RA, Gjedrem T (1970) Restriction of Selection Indexes. *Biometrics* 26:67. <https://doi.org/10.2307/2529045>
- Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics* 185:1021–1031. <https://doi.org/10.1534/genetics.110.116855>
- Dallinger HG, Löschenberger F, Bistrich H, et al (2023) Predictor bias in genomic and phenomic selection. *Theor Appl Genet* 136:235. <https://doi.org/10.1007/s00122-023-04479-8>
- Davies AMC (2005) An Introduction to near Infrared Spectroscopy. *NIR news* 16:9–11. <https://doi.org/10.1255/nirn.853>
- Degenhart DF, Kondra ZP (1984) Relationships between seed yield and growth characters, yield components, seed quality of summer-type oilseed rape (*Brassica napus* L.). *Euphytica* 33:885–889. <https://doi.org/10.1007/BF00021916>
- Difabachew YF, Frisch M, Langstroff AL, et al (2023) Genomic prediction with haplotype blocks in wheat. *Front Plant Sci* 14:1168547. <https://doi.org/10.3389/fpls.2023.1168547>
- Foley WJ, McIlwee A, Lawler I, et al (1998) Ecological applications of near infrared reflectance spectroscopy - a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. *Oecologia* 116:293–305. <https://doi.org/10.1007/s004420050591>
- Gebreselassie MN, Ader K, Boizot N, et al (2017) Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Industrial Crops and Products* 107:159–171. <https://doi.org/10.1016/j.indcrop.2017.05.013>
- González-Recio O, Rosa GJM, Gianola D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science* 166:217–231. <https://doi.org/10.1016/j.livsci.2014.05.036>
- Han X, Tang Q, Xu L, et al (2022) Genome-wide detection of genotype environment interactions for flowering time in *Brassica napus*. *Front Plant Sci* 13:1065766. <https://doi.org/10.3389/fpls.2022.1065766>
- Hart JR, Norris KH, Golumbic C (1962) Determination of the moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. *Cereal Chemistry* 39:94–99
- Henderson CR (1975) Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31:423. <https://doi.org/10.2307/2529430>
- Hocking RR, Pendleton OJ (1983) The regression dilemma. *Communications in Statistics - Theory and Methods* 12:497–527. <https://doi.org/10.1080/03610928308828477>

- Hu X, Carver BF, Powers C, et al (2019) Effectiveness of Genomic Selection by Response to Selection for Winter Wheat Variety Improvement. *The Plant Genome* 12:180090. <https://doi.org/10.3835/plantgenome2018.11.0090>
- Jiang HY, Zhu YJ, Wei LM, et al (2007) Analysis of protein, starch and oil content of single intact kernels by near infrared reflectance spectroscopy (NIRS) in maize (*Zea mays* L.). *Plant Breeding* 126:492–497. <https://doi.org/10.1111/j.1439-0523.2007.01338.x>
- Lane HM, Murray SC, Montesinos-López OA, et al (2020) Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. *The Plant Phenome Journal* 3:e20002. <https://doi.org/10.1002/ppj2.20002>
- Li F, Chen B, Xu K, et al (2016) A genome-wide association study of plant height and primary branch number in rapeseed (*Brassica napus*). *Plant Science* 242:169–177. <https://doi.org/10.1016/j.plantsci.2015.05.012>
- Li H, Lascano RJ, Barnes EM, et al (2001) Multispectral Reflectance of Cotton Related to Plant Growth, Soil Water and Texture, and Site Elevation. *Agronomy Journal* 93:1327–1337. <https://doi.org/10.2134/agronj2001.1327>
- Li N, Peng W, Shi J, et al (2015) The Natural Variation of Seed Weight Is Mainly Controlled by Maternal Genotype in Rapeseed (*Brassica napus* L.). *PLoS ONE* 10:e0125360. <https://doi.org/10.1371/journal.pone.0125360>
- Lin Y-C, Mayer M, Valle Torres D, et al (2024) Genomic prediction within and across maize landrace derived populations using haplotypes. *Front Plant Sci* 15:1351466. <https://doi.org/10.3389/fpls.2024.1351466>
- Lopez-Cruz M, Olson E, Rovere G, et al (2020) Regularized selection indices for breeding value prediction using hyper-spectral image data. *Sci Rep* 10:8195. <https://doi.org/10.1038/s41598-020-65011-2>
- Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120:151–161. <https://doi.org/10.1007/s00122-009-1166-3>
- Luan T, Woolliams JA, Lien S, et al (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* 183:1119–1126. <https://doi.org/10.1534/genetics.109.107391>
- McVetty PBE, Pinnisch R (1994) Comparison of the effect of *nap* and *pol* cytoplasm on the performance of three summer oilseed rape cultivar-derived isoline pairs. *Can J Plant Sci* 74:729–731. <https://doi.org/10.4141/cjps94-130>
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157:1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>

- Nowosad K, Liersch A, Poptawska W, Bocianowski J (2016) Genotype by environment interaction for seed yield in rapeseed (*Brassica napus* L.) using additive main effects and multiplicative interaction model. *Euphytica* 208:187–194. <https://doi.org/10.1007/s10681-015-1620-z>
- Ou J-H, Liao C-T (2019) Training set determination for genomic selection. *Theor Appl Genet* 132:2781–2792. <https://doi.org/10.1007/s00122-019-03387-0>
- Pearson K, Lee A (1903) On the Laws of Inheritance in Man: I. Inheritance of Physical Characters. *Biometrika* 2:357. <https://doi.org/10.2307/2331507>
- Piepho H-P, Möhring J (2007) Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials. *Genetics* 177:1881–1888. <https://doi.org/10.1534/genetics.107.074229>
- Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161:209–228. <https://doi.org/10.1007/s10681-007-9449-8>
- Pojić M, Mastilović J, Majcen N (2012) Robustness of the near infrared spectroscopy method determined using univariate and multivariate approach. *Food Chemistry* 134:1699–1705. <https://doi.org/10.1016/j.foodchem.2012.03.104>
- Pritchard FM, Eagles HA, Norton RM, et al (2000) Environmental effects on seed composition of Victorian canola. *Aust J Exp Agric* 40:679. <https://doi.org/10.1071/EA99146>
- Rincent R, Charpentier J-P, Faivre-Rampant P, et al (2018) Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3 Genes|Genomes|Genetics* 8:3961–3972. <https://doi.org/10.1534/g3.118.200760>
- Robert P, Auzanneau J, Goudemand E, et al (2022a) Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theor Appl Genet* 135:895–914. <https://doi.org/10.1007/s00122-021-04005-8>
- Robert P, Brault C, Rincent R, Segura V (2022b) Phenomic Selection: A New and Efficient Alternative to Genomic Selection. In: Ahmadi N, Bartholomé J (eds) *Genomic Prediction of Complex Traits*. Springer US, New York, NY, pp 397–420
- Roscher-Ehrig L, Weber SE, Abbadi A, et al (2024) Phenomic Selection for Hybrid Rapeseed Breeding. *Plant Phenomics* 6:0215. <https://doi.org/10.34133/plantphenomics.0215>
- Savitzky Abraham, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* 36:1627–1639. <https://doi.org/10.1021/ac60214a047>

- Schiesl S, Iniguez-Luy F, Qian W, Snowdon RJ (2015) Diverse regulatory factors associate with flowering time and yield responses in winter-type *Brassica napus*. *BMC Genomics* 16:737. <https://doi.org/10.1186/s12864-015-1950-1>
- Schmidt P, Hartung J, Bennewitz J, Piepho H-P (2019) Heritability in Plant Breeding on a Genotype-Difference Basis. *Genetics* 212:991–1008. <https://doi.org/10.1534/genetics.119.302134>
- Seregély Z, Deák T, Bisztray GD (2004) Distinguishing melon genotypes using NIR spectroscopy. *Chemometrics and Intelligent Laboratory Systems* 72:195–203. <https://doi.org/10.1016/j.chemolab.2004.01.013>
- Stevens A, Ramirez-Lopez L (2025) An introduction to the prospectr package
- Sweeney DW, Rooney TE, Sorrells ME (2021) Gain from genomic selection for a selection index in two-row spring barley. *The Plant Genome* 14:e20138. <https://doi.org/10.1002/tpg2.20138>
- Tkachuk R (1981) Oil and protein analysis of whole rapeseed kernels by near infrared reflectance spectroscopy. *J Americ Oil Chem Soc* 58:819–822. <https://doi.org/10.1007/BF02665588>
- VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>
- Variath MT, Wu JG, Li YX, et al (2009) Genetic analysis for oil and protein contents of rapeseed (*Brassica napus* L.) at different developmental times. *Euphytica* 166:145–153. <https://doi.org/10.1007/s10681-008-9851-x>
- Wang X, Liu G, Yang Q, et al (2010) Genetic analysis on oil content in rapeseed (*Brassica napus* L.). *Euphytica* 173:17–24. <https://doi.org/10.1007/s10681-009-0062-x>
- Weber SE, Frisch M, Snowdon RJ, Voss-Fels KP (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Front Plant Sci* 14:1217589. <https://doi.org/10.3389/fpls.2023.1217589>
- Weir BS (1979) Inferences about Linkage Disequilibrium. *Biometrics* 35:235. <https://doi.org/10.2307/2529947>
- Werteker M, Lorenz A, Johannes H, et al (2010) Environmental and Varietal Influences on the Fatty Acid Composition of Rapeseed, Soybeans and Sunflowers. *J Agronomy Crop Science* 196:20–27. <https://doi.org/10.1111/j.1439-037X.2009.00393.x>
- Williams PC, Mackenzie SL, Starkey PM (1985) Determination of methionine in peas by near-infrared reflectance spectroscopy (NIRS). *J Agric Food Chem* 33:811–815. <https://doi.org/10.1021/jf00065a011>
- Won S, Park J-E, Son J-H, et al (2020) Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium. *Front Genet* 11:134. <https://doi.org/10.3389/fgene.2020.00134>

- Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (*B. rassica napus* L.) breeding. Plant Breeding 133:45–51.
<https://doi.org/10.1111/pbr.12137>
- Zaykin DV, Pudovkin A, Weir BS (2008) Correlation-Based Inference for Linkage Disequilibrium With Multiple Alleles. Genetics 180:533–545.
<https://doi.org/10.1534/genetics.108.089409>
- Zhu X, Leiser WL, Hahn V, Würschum T (2021) Phenomic selection is competitive with genomic selection for breeding of complex traits. The Plant Phenome Journal 4:e20027. <https://doi.org/10.1002/ppj2.20027>
- Zhu X, Maurer HP, Jenz M, et al (2022) The performance of phenomic selection depends on the genetic architecture of the target trait. Theor Appl Genet 135:653–665.
<https://doi.org/10.1007/s00122-021-03997-7>
- Zimmermann B, Kohler A (2013) Optimizing Savitzky–Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy. Appl Spectrosc 67:892–902. <https://doi.org/10.1366/12-06723>

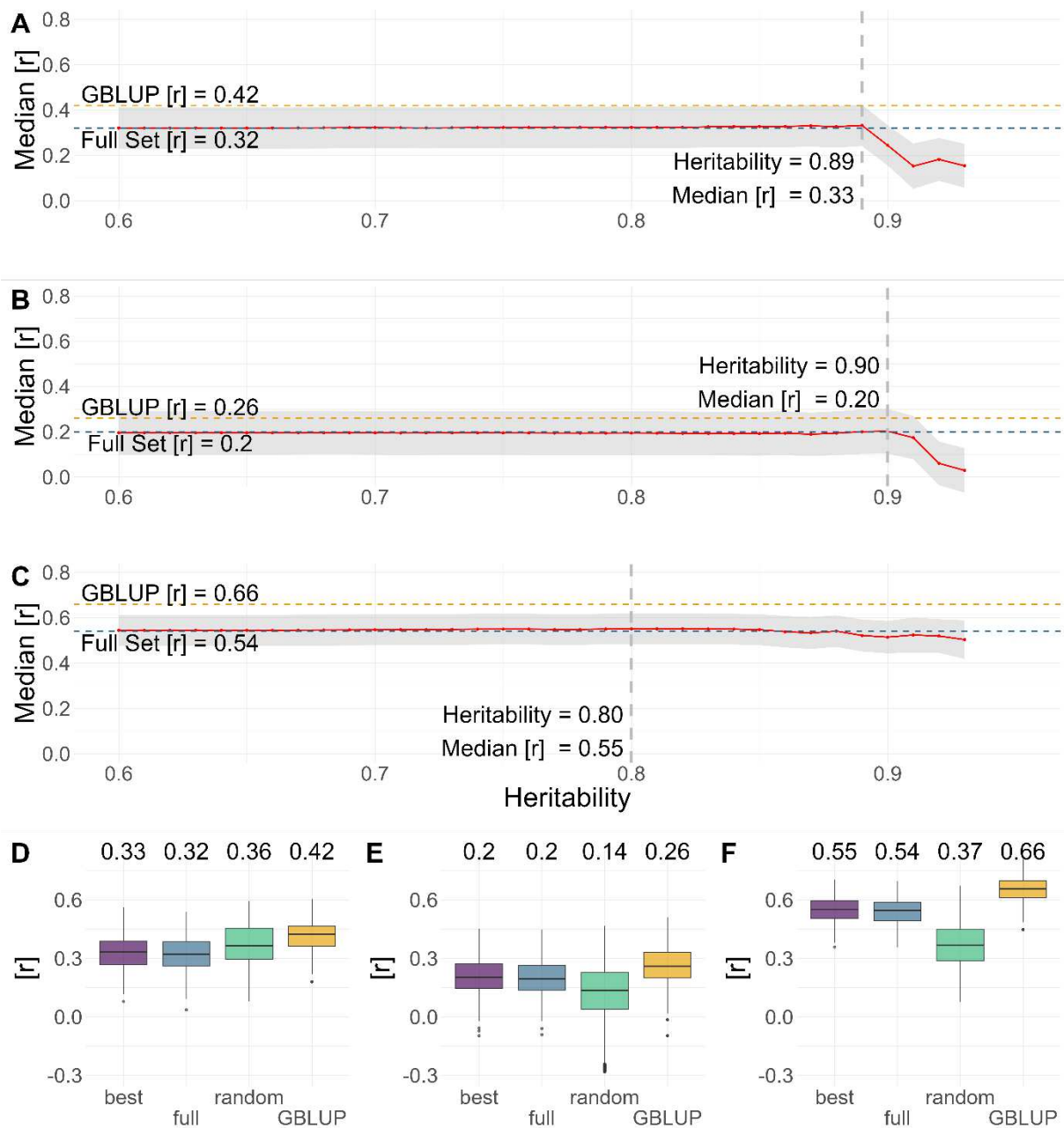


Fig. S1 Across-generation prediction of (A; D) seed yield, (B;E) plant height and (C;F) flowering time after heritability filtering. The grey dashed line indicates the best subset along with its accuracy and the applied heritability threshold. The blue dashed line represents the accuracy achieved using the full spectral dataset, while the orange dashed line shows the prediction accuracy of the corresponding GBLUP model. D, E and F show the median prediction accuracy of the best subset, the full set, the random subset and the GBLUP model with its standard deviation, respectively.

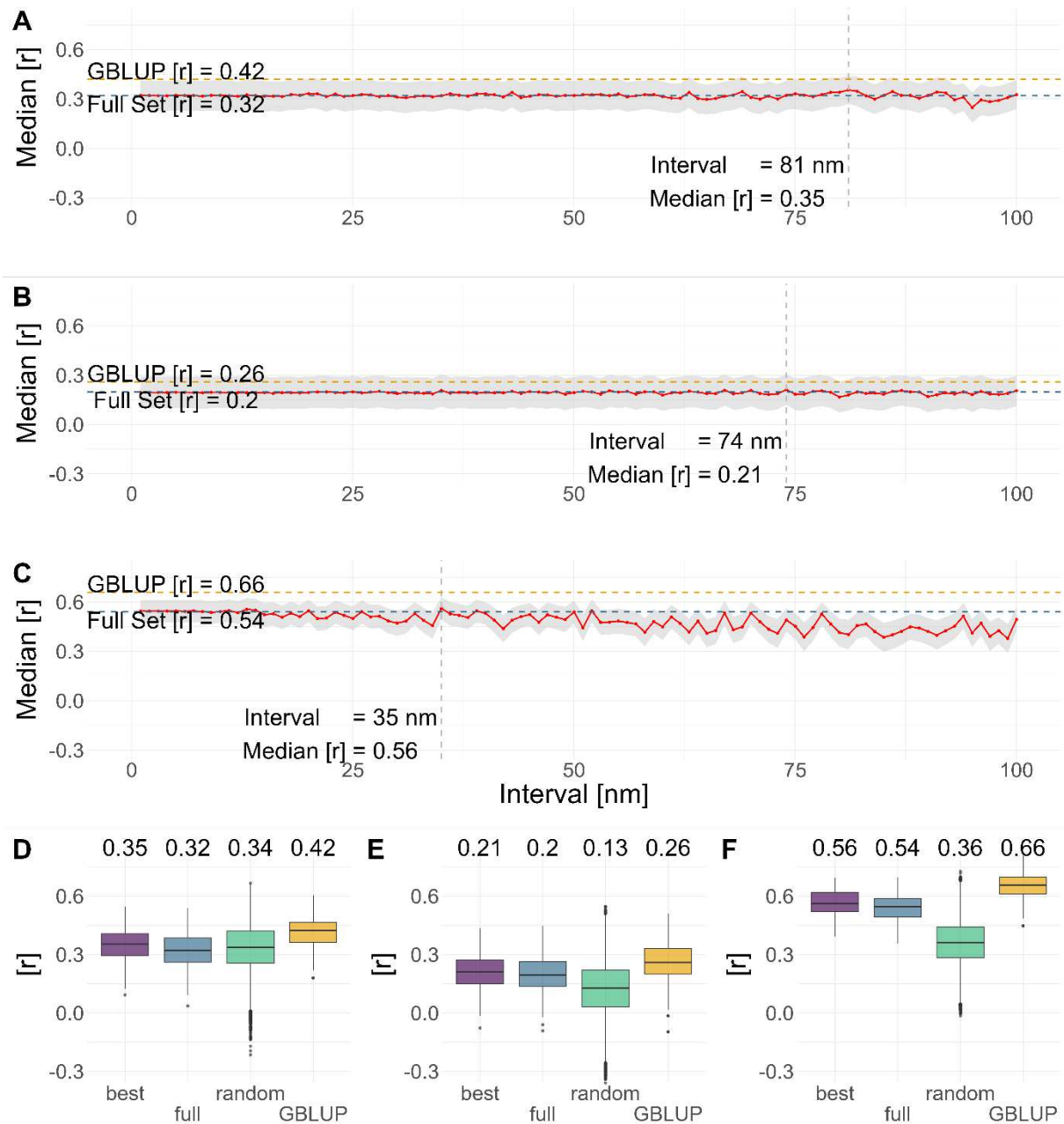


Fig. S2 Across-generation prediction of (A; D) seed yield, (B;E) plant height and (C;F) flowering time after resolution filtering. The grey dashed line indicates the best subset along with its accuracy and the interval between the wavelengths. The blue dashed line represents the accuracy achieved using the full spectral dataset, while the orange dashed line shows the prediction accuracy of the corresponding GBLUP model. D, E and F show the median prediction accuracy of the best subset, the full set, the random subset and the GBLUP model with its standard deviation, respectively.

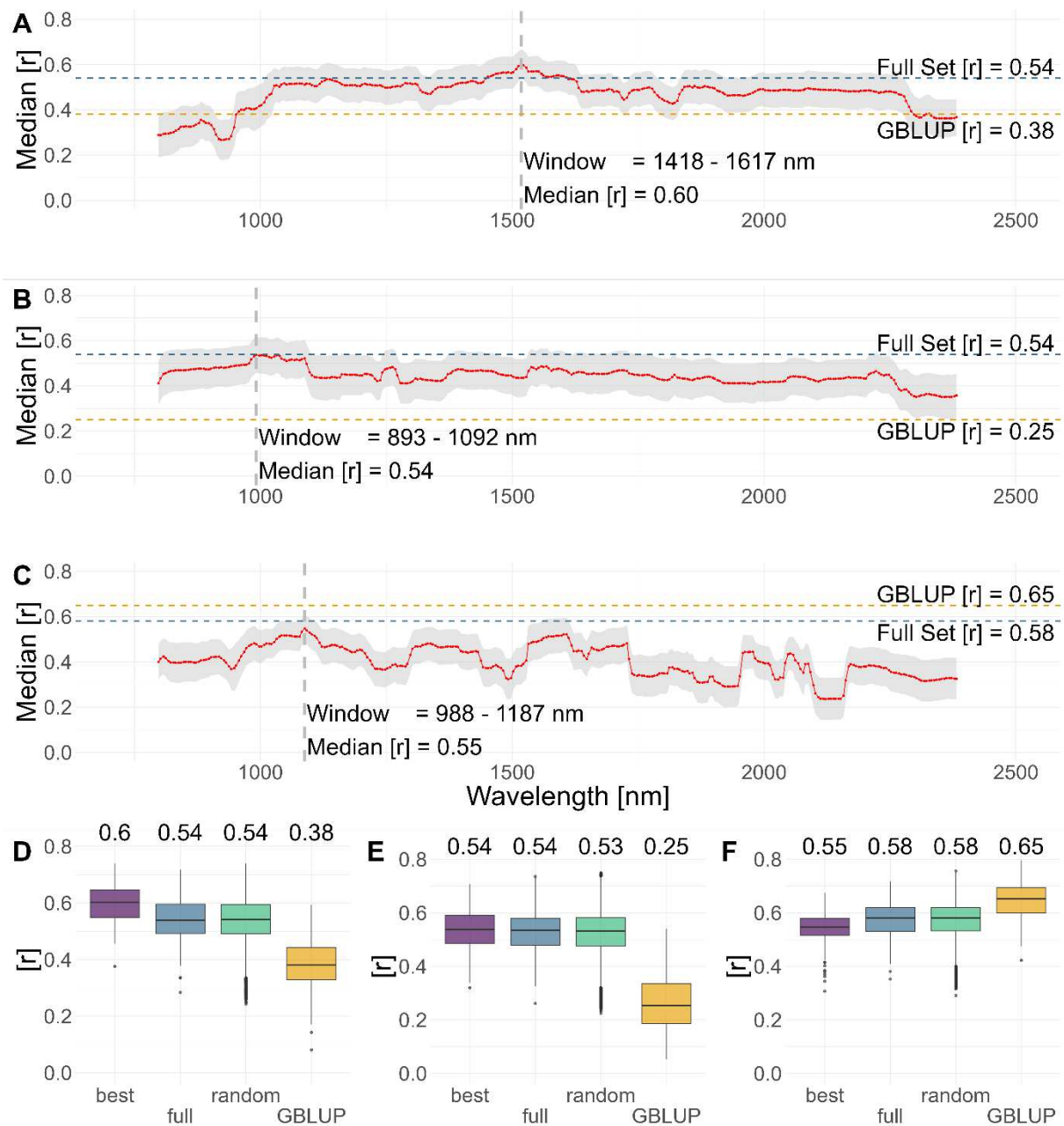


Fig. S3 Median prediction accuracy of within-generational predictions. Each red dot represents the median prediction accuracy of a single 200 nm window at the mean wavelength of the window for the traits seed yield(A), plant height (B), and flowering time (C). The grey dashed line indicates the best subset, while the full set accuracy and GBLUP accuracy are indicated by the blue and yellow lines (A–C), respectively. Boxplots on the right-hand side (D–F) show the corresponding best, full, random, and GBLUP sets. Bars represent the standard deviation, and lines indicate the median prediction accuracy (values given above) obtained after 200 fold cross validation.

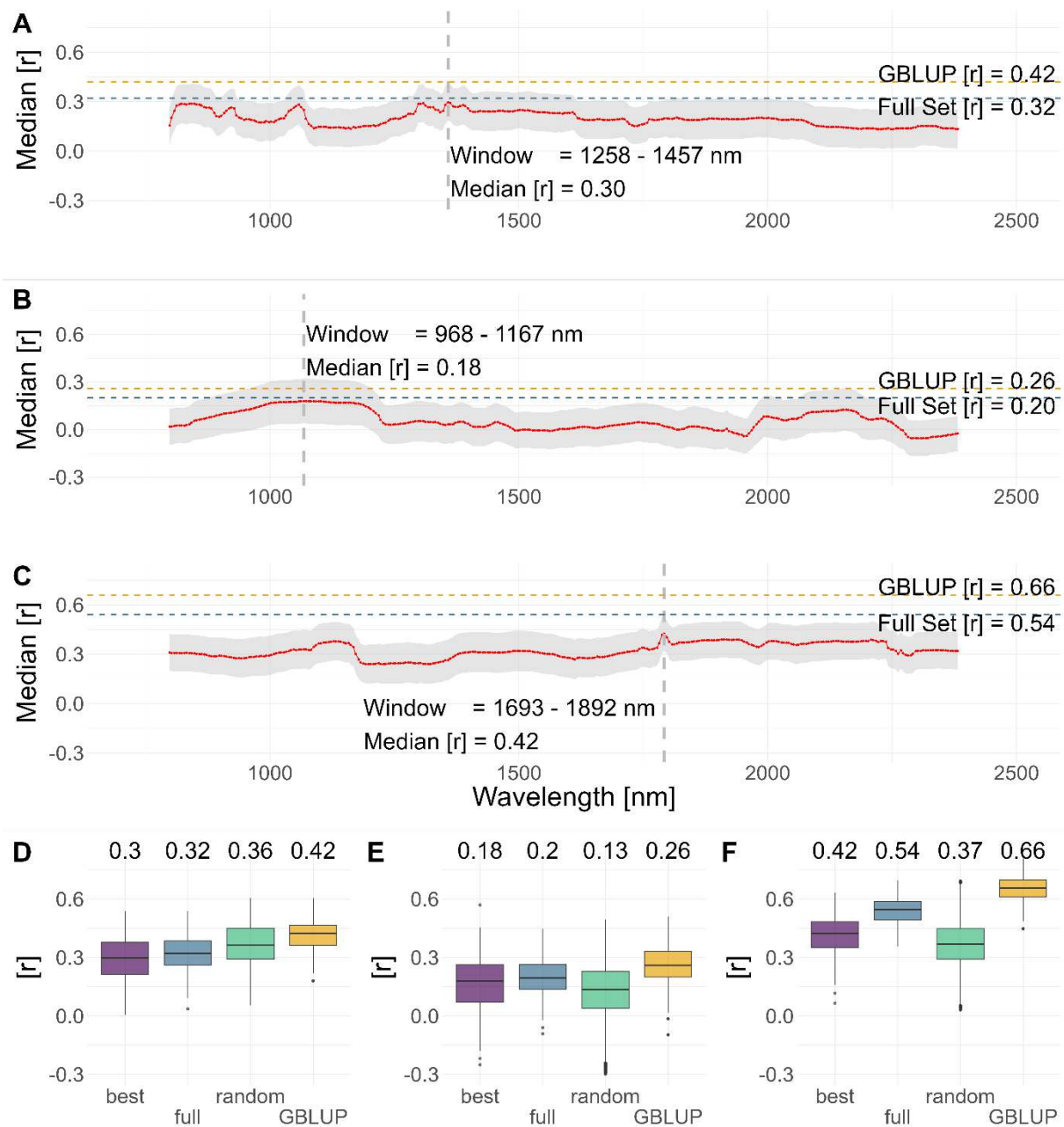


Fig. S4 Median prediction accuracy of across-generational predictions. Each red dot represents the median prediction accuracy of a single 200 nm window at the mean wavelength of the window for the traits seed yield(A), plant height (B), and flowering time (C). The grey dashed line indicates the best subset, while the full set accuracy and GBLUP accuracy are indicated by the blue and yellow lines (A–C), respectively. Boxplots on the right-hand side (D–F) show the corresponding best, full, random, and GBLUP sets. Bars represent the standard deviation, and lines indicate the median prediction accuracy (values given above) obtained after 200 fold cross validation.