



Simultaneous estimation of yield and protein content in barley using derived traits and near-infrared spectroscopy

B. Dyulgerova

Agricultural Academy, Institute of Agriculture - Karnobat, 8400 Karnobat, Bulgaria

(Manuscript received 29 April 2026; accepted for publication 26 August 2026)

Abstract. This study aimed to establish relationships among grain yield, protein content, and derived traits and indices, as well as the possibility of using near-infrared (NIR) spectroscopy for simultaneous selection of grain yield and protein content in barley breeding. A total of 50 genotypes of winter six-row barley were included in the study, consisting of 23 mutant lines derived from the Monika variety, 24 from the Izgrev variety, their parental varieties, and the standard control variety Veslets. Field experiments were conducted over three growing seasons (2021/2022-2023/2024) in the experimental field of the Institute of Agriculture-Karnobat, Southeastern Bulgaria. NIR spectra were acquired with a CropScan 1000B Near-Infrared Analyzer over the 720–1100 nm range at 10 nm intervals, and predictive models were built in R (v4.4.2) using a gradient boosting algorithm (XGBoost) with a pairwise ranking objective, evaluated through 5-fold cross-validation. Analysis of variance revealed significant effects of year, genotype, and their interaction on grain yield, protein content, and protein yield. The largest share of the variation in these traits was due to the conditions of the growing year. A strong negative correlation between yield and protein content was observed in the studied mutant lines. Grain protein deviation (GPD) was the most informative derived trait, reducing the dependence on yield and allowing more efficient identification of high-protein genotypes. In contrast, PY, GP&YD and EW showed strong redundancy with primary traits, resulting in minimal added information value. Predictive ability, assessed as the Spearman rank correlation between predicted and observed values, was strongly trait-dependent: protein content was predicted with high accuracy ($R \approx 0.90-0.95$), grain yield with moderate accuracy ($R \approx 0.19-0.56$), and GPD with intermediate but variable accuracy across years ($R \approx 0.40-0.80$), while the remaining composite indices showed weak and unstable predictive ability. The predictive performance based on NIR was highly dependent on the trait. Protein content showed high predictability, while grain yield demonstrated moderate predictive performance. GPD showed intermediate but variable predictability across years, while other composite indices were poorly predicted, likely due to poor spectral representation and indirect biological significance. NIR spectroscopy can be very useful for the rapid selection of grain quality traits, especially protein content, while its utility for yield and complex indices is limited and requires additional complementary data.

Keywords: barley, grain yield, protein content, protein yield, NIR spectroscopy

Introduction

The main part of winter barley grain is used for animal feed, and breeding has traditionally focused on increasing grain yield. However, grain protein content is also an important determinant of feed quality, making the simultaneous improvement of productivity and protein concentration an important breeding objective. This is challenging because grain yield and protein content are generally negatively correlated, limiting the efficiency of direct selection for both traits simultaneously (Polisetty and Chandra, 1989; Simmonds, 1995). Various derived traits and selection indices have therefore been proposed to integrate yield and protein into a single selection criterion. Koekemoer et al. (1999) proposed protein yield, which represents the amount of protein harvested per unit area and enables the identification of genotypes combining high grain productivity with desirable protein levels. Another approach is

grain protein deviation (GPD), which accounts for the deviation of observed grain protein content from that expected based on grain yield (Monaghan et al., 2001). Studies in bread (*T. aestivum* ssp. *aestivum*) and durum (*T. turgidum* ssp. *durum*) wheat have demonstrated that GPD is under genetic control and can therefore be exploited for selection (Rapp et al., 2018; Nigro et al., 2019; Mosleth et al., 2015; Geyer et al., 2022; Paina and Gregersen, 2023; Richard et al., 2025). Similarly, the GP&YD and EW indices proposed by Rapp et al. (2018) combine standardized components of grain protein and yield, providing more balanced selection criteria for simultaneous improvement of the two traits.

Near-infrared (NIR) spectroscopy provides a rapid, non-destructive and cost-effective approach for evaluating grain quality traits and is widely used for determining protein content in barley (Caporaso et al., 2018). More recently, its application has expanded beyond conventional analytical purposes to

*e-mail: bdyulgerova@abv.bg

phenotypic breeding, where spectral information is used as a predictor of complex traits, including grain yield (Robert et al., 2022; Setotaw et al., 2024; Bienvenu et al., 2025). Because NIR spectra can capture information associated with multiple grain properties, the approach may provide an opportunity to assess yield- and quality-related traits within the same analytical framework and thereby support simultaneous selection for productivity and grain protein content. Based on this premise, we hypothesized that NIR spectroscopy could be used to facilitate simultaneous selection for grain yield and protein content in barley breeding. The aim of this study was therefore to investigate the relationships among grain yield, protein content and related derived traits and selection indices, and to evaluate the potential of NIR spectroscopy for their simultaneous assessment and selection in barley breeding.

Material and methods

Plant material and experimental design

The study included 50 winter six-rowed barley genotypes, consisting of 23 mutant lines derived from variety Monika, 24 from variety Izgrev, their parental varieties, and the standard check variety Veslets.

Field experiments were conducted over three growing years (2021/2022-2023/2024) at the experimental field of the Institute of Agriculture, Karnobat, South-eastern Bulgaria, using a complete block design with four replications per genotype.

Traits evaluation

Grain yield (GY) and grain protein content (PC) were recorded for all genotypes in three years. Grain yield (kg ha⁻¹) was measured as grain production per unit area. Protein content (%) was determined by the Kjeldahl method.

Calculation of indices

Protein yield (PY, kg ha⁻¹) was calculated on plot basis as the product of GY and PC and represents the amount of harvested protein in kilograms per ha.

Grain protein deviation (GPD) was calculated using the bivariate linear mixed model approach suggested by Thorwarth et al. (2018) to account for covariance effects of grain yield and protein content due to the experimental design.

Best linear unbiased predictors (BLUPs) of genotypic effects for grain yield and protein content were obtained using a multivariate mixed model. The genetic variance-covariance matrix was estimated, and the genetic correlation between traits was calculated.

Grain protein deviation was computed as the residual of protein content after removing its linear dependency on grain yield:

$$GPD = Y_{PC} - \rho \frac{\sigma_{PC}}{\sigma_{GY}} Y_{GY}$$

where Y_{PC} and Y_{GY} represent the trait values of PC and GY, ρ represents the genetic correlation between PC and GY and σ_{PC} and σ_{GY} denote the standard deviation of PC and GY.

Index GP&YD and *Index EW* were suggested by Rapp et al. (2018), where Index GP&YD represents the sum of standardized

GPD and YD and Index EW is the sum of standardized PC and GY. Standardization was done according to formula (2):

$$Z = \frac{x - \hat{\mu}}{\sigma}$$

where Z represents the obtained standardized trait value, X denotes the plot value, $\hat{\mu}$ –the overall mean and σ the standard deviation of the respective trait.

Spectral data acquisition and preprocessing

Near-infrared (NIR) spectral data were collected using a CropScan 1000B Near-Infrared Analyzer (Next Instruments Pty Ltd, Condell Park, NSW 2200, Australia), covering the 720-1100 nm wavelength range at 10 nm intervals, yielding 38 spectral variables per sample. Bulk whole-grain samples (500 g) from each plot were scanned in reflectance mode. Each sample was scanned 15 times, and the average spectrum was used to represent the plot's spectral profile.

NIR spectra were preprocessed using Savitzky–Golay second-derivative transformation to reduce noise and correct for scatter effects (Savitzky and Golay, 1964).

Predictive modelling

A gradient boosting algorithm implemented in the XGBoost framework was applied using a pairwise ranking objective (rank:pairwise). This approach models the relative ordering of genotypes rather than absolute trait values and is therefore suitable for selection-oriented analysis.

Model validation

Predictive performance was evaluated using a repeated 5-fold cross-validation scheme, applied independently within each growing year to account for the strong year effect identified in the analysis of variance. In each repetition, the dataset was randomly divided into training (4 folds) and testing (1 fold) subsets, until each fold served once as the test set. The random partitioning into folds was repeated multiple times to reduce the influence of a single, potentially unrepresentative data split on the estimated predictive ability. Within each training set, all genotypes were treated as a single ranking group, as required by the pairwise ranking formulation, meaning that the model learned to order genotypes relative to one another within that group rather than to predict their individual trait values independently. This cross-validation strategy provides an internal, resampling-based estimate of model transferability by repeatedly assessing performance on genotypes not included in model training, which is a standard approach for evaluating predictive ability in phenomic and genomic selection studies (Robert et al., 2022; Setotaw et al., 2024).

Model parametrization

To minimize overfitting, given the limited sample size relative to model complexity, conservative hyperparameters were used rather than values obtained through exhaustive tuning: learning rate (eta)=0.05, maximum tree depth=3, subsample ratio=0.7, column subsample ratio=0.7, and minimum child weight=5. These values were fixed a priori based on common practice reported for gradient boosting applications to small-to-moderate sized breeding

datasets (Chen and Guestrin, 2016; Robert et al., 2022), avoiding the additional risk of overfitting associated with hyperparameter optimization on a limited number of genotypes. Shallow trees (depth=3) restrict the complexity of individual base learners, while row and column subsampling introduce stochasticity that reduces variance and the risk of the model fitting noise specific to the training folds. Models were trained for up to 500 boosting iterations, with early stopping applied after 20 consecutive rounds without improvement in validation performance, which prevents unnecessary tree addition once predictive gains plateau and further limits overfitting. All models were implemented using the xgboost R package (version 1.7.8.1).

Performance assessment

Predictive ability was evaluated using Spearman's rank correlation coefficient between predicted and observed values in the test sets, consistent with the ranking-based nature of the model. Both predicted and observed values were converted to ranks prior to correlation analysis. Final model performance was expressed as the mean and standard deviation of Spearman correlations across cross-validation folds within each year, allowing the assessment of both the overall predictive ability and its consistency across independent growing seasons. The ranking-based approach focuses on identifying superior genotypes rather than predicting exact trait values, which makes it particularly suitable for phenomic selection, where relative performance is more important than absolute accuracy (Robert et al., 2022).

All statistical analyses were performed using the R environment (version 4.4.2).

Results and discussion

Variation in grain yield, protein content, and protein yield across genotypes and years

The analysis of variance showed statistically significant effects of year (Y), genotype (G) and their interaction (G×Y) on grain yield (GY), protein content (PC) and protein yield (PY) (Table 1).

For all studied traits, year (Y) was the main source of variation, indicating the dominant importance of agro-ecological conditions for the phenotypic expression of these traits. The effect of the growing year explained 72.76% of the total variation in grain yield, 74.26% in protein content and 79.41% in protein yield.

The effect of genotype (G) was smaller than that of the growing year. The proportion of variance explained by this factor was 18.86% for grain yield, 11.52% for protein content, and 10.08% for protein yield. The significant genotype effect indicates the presence of substantial genetic variability among the tested mutant lines for the studied traits, which provides an opportunity for effective selection.

The contribution of the genotype × year (G×Y) interaction to the total variation was the lowest for grain yield (7.87%) followed by protein yield (9.88%) and protein content (13.67%).

Table 1. Mean squares and percentage contribution of sum of squares, % for grain yield (GY), protein content (PC), and protein yield (PY) of 50 barley genotypes across three growing seasons

Traits	Mean squares			Sum of squares, %		
	G	Y	G×Y	G	Y	G×Y
GY	206046297.85*	152.84*	1819864.95*	18.86	72.76	7.87
PC	794784081.72*	984.75*	14337553.92*	11.52	74.26	13.67
PY	85995360.28*	181.25*	1783915.56*	10.08	79.41	9.88

*indicates significance at $p < 0.001$

The mean, minimum and maximum values, and the coefficient of variation (CV%) for grain yield (GY), protein content (PC) and protein yield (PY) for the three growing years, are presented in Table 2.

The results confirmed the significant variability between years for the studied traits, which was identified by the analysis of variance. The highest mean grain yield (GY) was recorded in 2023 (7345.50 kg ha⁻¹) and the lowest in 2024 (4539.66 kg ha⁻¹). The difference in yield between the two years was nearly 40%, showing the strong effect of year-to-year conditions on the trait. The coefficient of variation (CV) for GY remained relatively stable between years (11.47-12.36%), suggesting similar variability despite large differences in mean values.

Protein content (PC) had the highest mean value in 2022 (13.99%) and the lowest in 2024 (11.02%). PC showed lower variability compared to grain yield, with CV ranging from 4.79% to 7.95%.

Protein yield (PY) followed grain yield trends, indicating a strong dependence of this trait on grain yield. The highest average value was recorded in 2023 (849.51 kg ha⁻¹), while the lowest was in 2024 (498.40 kg ha⁻¹). The coefficient of variation values for PY (9.86-11.71%) were in a similar range to those for GY.

The overall data for the three years showed significantly lower variability in protein content compared to grain yield, suggesting that the studied mutant lines differed more in terms of productivity than in terms of the quality trait – protein.

Table 2. Descriptive statistics for grain yield (GY, kg ha⁻¹), protein content (PC, %), and protein yield (PY, kg ha⁻¹) of 50 barley genotypes across three growing seasons. Overall statistics were calculated based on genotype means averaged across years.

Year	Statistic	GY	PC	PY
2022	Mean	5705.20	13.99	796.70
	Min	4537.50	12.36	616.19
	Max	7435.00	15.11	968.60
	CV (%)	11.96	5.41	11.71
2023	Mean	7345.50	11.62	849.51
	Min	5060.00	10.14	663.75
	Max	8749.00	14.39	1024.95
	CV (%)	11.47	7.95	9.86
2024	Mean	4539.66	11.02	498.40
	Min	3336.25	9.45	385.50
	Max	5627.50	11.97	615.81
	CV (%)	12.36	4.79	10.36
Overall	Mean	5863.45	12.21	714.87
	Min	4548.75	10.95	576.17
	Max	6983.75	13.55	823.18
	CV (%)	10.10	4.18	7.78

Relationships among grain yield, grain protein content and derived traits

Figure 1 presents the results of the correlations between grain yield, grain protein content and derived traits. Grain yield showed strong positive correlations with protein yield (PY, $r=0.91$) and the Index GP&YD ($r=0.81$). The very strong association between PY and the Index GP&YD ($r=0.98$) suggests a considerable overlap in the information captured by these variables in relation to protein yield per unit area. A strong negative correlation was found between grain protein content and grain yield ($r=-0.70$), confirming the well-known yield-protein trade-off in cereals. Therefore, selecting breeding lines for higher yield alone is likely to reduce grain protein concentration, unless breeding strategies are specifically designed to balance this trade-off. Remarkably, the negative correlation between yield and protein content was maintained in the mutant population analyzed in this study, suggesting that induced mutagenesis did not affect significantly the established physiological relationship between these two traits.

GPD showed a moderate positive correlation with PY ($r=0.64$) and no relationship with GY ($r=-0.18$), indicating that GPD effectively removes the direct dependence on yield while retaining some of the genetic signal associated with protein. Previous studies have highlighted the value of GPD for selecting varieties producing higher protein concentrations than expected based on their yield (Monaghan et al., 2001; Oury and Godin, 2007; Paina and Gregersen, 2023). The observed strong positive relationship between GPD and the standardized Index EW ($r=0.82$) indicates that both indices describe closely related aspects of protein-related variation.

The Index GP&YD and the Index EW integrated yield and protein content in different ways. The Index EW showed strong positive relationships with PY ($r=0.86$) and GY ($r=0.60$), which suggests that the index was more closely related to overall productivity than to protein efficiency. In contrast, the Index GP&YD was correlated with GY ($r=0.41$) and was not correlated with PC ($r=-0.17$), suggesting that this index reflected a moderate degree of variation in yield but not in protein concentration. This lack of relationship with PC implies that the Index GP&YD probably will be less appropriate when the goal is to simultaneously improve both yield and protein content. Instead, it appears to emphasize yield-related traits without adequately considering protein quality, which could limit its usefulness in barley breeding programs.

The results confirm the antagonism between grain yield and protein content, which remains a major constraint in cereal breeding. While PY integrates both traits into a single measure of total protein production, it does not resolve the fundamental trade-off between quantity and quality. GPD, derived from a mixed model framework accounting for genetic covariance, provides a more precise measure of protein productivity independent of yield and is therefore particularly valuable for selection of genotypes that violate the typical yield-protein trade-off.

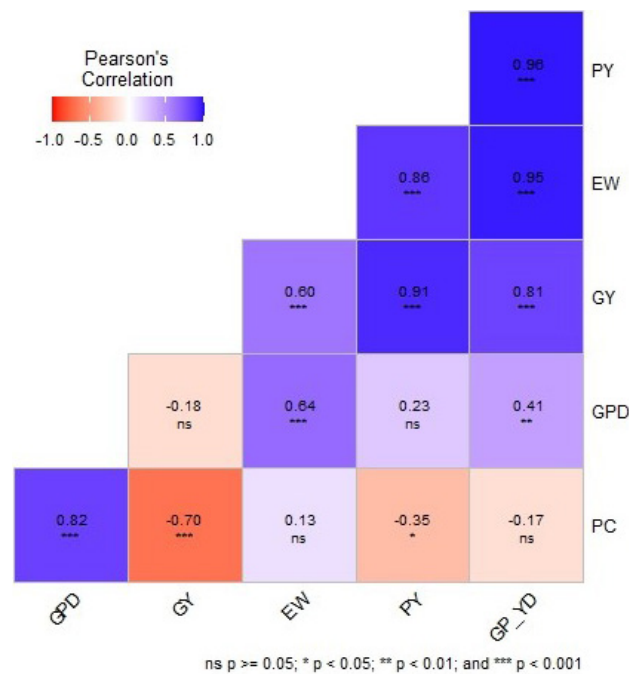


Figure 1. Pearson correlation coefficients among grain yield (GY), protein content (PC), protein yield (PY), grain protein deviation (GPD), Index GP&YD (GP_YD), and Index EW (EW).

Genotype Selection across Traits and Indices

Relatively low overlap between genotypes selected for individual traits and indices at the 20% selection intensity was observed (Figure 2). This clearly demonstrates the presence of significant rearrangement of genotypes depending on the selection criterion used. Low correspondence was found between selection for yield (GY) and protein content (PC), confirming the well-known trade-off between the two traits. While, the use of derived traits and integrative indices, such as PY, Index GP&YD and Index EW, led to partial overlap of the selected genotypes, which indicates their ability to combine information from both traits. However, the differences between these indices indicate that they cover different aspects of the interaction between yield and protein content. This suggests that the choice of selection index plays a critical role in genotype selection and must be chosen in line with the specific goals of the breeding program.

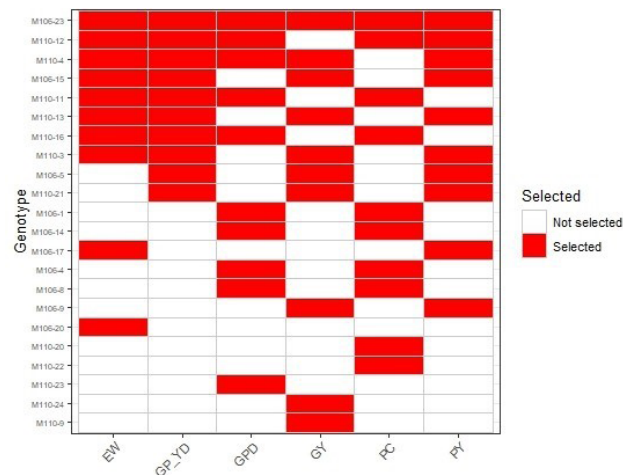


Figure 2. Selected genotypes based on grain yield (GY), protein content (PC), protein yield (PY), grain protein deviation (GPD), Index GP&YD (GP_YD) and Index EW (EW) (20% selection intensity)

Prediction of grain yield, protein content, and derived traits using NIR spectroscopy

Protein content (PC) was predicted with consistently very high accuracy ($R \approx 0.90-0.95$) and low standard deviation (Table 3). This is probably directly related to the fact that NIR spectra capture nitrogen-related absorption features associated with proteins in the grain. The stability of the prediction over the years further confirms the stable and repeatable spectral signals for PC.

In contrast, grain yield (GY) was found to have moderate predictive power ($R \approx 0.19-0.56$). The lower accuracy compared to PC is likely consistent with the complex, polygenic, and environmentally dependent nature of yield, as well as its weaker direct relationship with spectral features.

Grain protein deviation (GPD), which was independent of yield and moderately related to protein variation, demonstrated moderate to very good predictive power ($R \approx 0.40-0.80$). This implies that GPD partially preserved spectral information related to protein content.

PY, Index GP&YD and Index EW, which were strongly dependent on yield, showed low and unstable predictive power. In particular, for PY and Index GP&YD, near-zero or even negative correlations were found, indicating poor model performance in genotype ranking.

The strong relationships identified between GY, PY, and index-based traits, combined with the high redundancy between them, are most likely the reason why the NIR model cannot

extract additional predictive information from the derived traits beyond what is already captured by the primary traits.

In contrast, GPD contributes additional information, which explains its relatively better performance compared to other derived indices.

The results show that the predictive capacity of the NIR-based model is largely determined by trait-specific characteristics. Traits with a direct chemical basis, such as protein content, were predicted with high accuracy as demonstrated in previous studies (Fox et al., 2002; Gous et al., 2012). More complex traits influenced by complicated physiological processes and interactions with the environment, such as grain yield, were predicted with moderate success, in agreement with findings reported by Robert et al. (2022) Berisha et al. (2025), Braun et al. (2025). Composite traits and indices were poorly predicted, as their indirect biological interpretation, inherent error propagation, and weak spectral representation limit the ability of models to capture reliable signals.

From a breeding perspective, these findings indicate a practical strategy in which NIR spectroscopy can be effectively exploited for rapid and cost-efficient selection of grain quality traits, particularly protein content, while its role in yield prediction remains supportive. For complex traits or indices, integration with additional data sources, such as environmental covariates or genomic information, is likely necessary to achieve reliable predictive performance.

Table 3. Predictive ability ($R \pm SD$) of the XGBoost model based on NIR data for different traits across three years

Year	GY	PC	PY	GPD	GP_YD	EW
2022	0.19 ± 0.14	0.93 ± 0.09	-0.02 ± 0.06	0.79 ± 0.16	0.23 ± 0.27	0.29 ± 0.23
2023	0.42 ± 0.52	0.95 ± 0.04	0.10 ± 0.31	0.40 ± 0.34	0.11 ± 0.38	0.28 ± 0.16
2024	0.56 ± 0.11	0.90 ± 0.07	0.39 ± 0.18	0.48 ± 0.19	0.03 ± 0.45	0.26 ± 0.28

GY - grain yield, PC - protein content, PY - protein yield, GPD - grain protein deviation, GP_YD - Index GP&YD, and EW - Index EW,

Conclusion

The observed variation in grain yield, protein content, and protein yield was largely determined by environmental conditions, while significant genotype effects indicated useful genetic variability. A strong negative correlation between yield and protein content was observed in the set of studied mutant lines. Grain protein deviation (GPD) was the most informative derived trait, as it minimized yield dependence and improved the identification of genotypes with superior protein performance. In contrast, other indices demonstrated strong redundancy with primary traits, resulting in minimal additional informational value. The predictive ability of NIR-based models was strongly trait-dependent. High predictive accuracy was obtained for protein content, whereas grain yield showed only moderate predictability. GPD showed intermediate but variable predictive ability across years. The remaining composite traits were weakly predicted, reflecting their indirect and complex biological nature.

Acknowledgements

This work was supported by the Bulgarian National Science Fund of the Ministry of Education and Science under project №

KP-06-COST-1/21.05.2024, as part of the national co-financing for Bulgarian participation in the European COST program (PANGEOS COST Action CA22136).

References

- Berisha P, Michel S, Garcia M, Michel M, Dutriez S and Bürstmayr H, 2025. Forward prediction of grain yield across early and advanced generations in winter wheat (*Triticum aestivum* L.) using NIRS-based models. The Plant Phenome Journal, 8, e70049. doi: <http://dx.doi.org/10.1002/ppj2.70049>
- Bienvenu C, Garin V, Salas N, Théra K, Tekete ML, Sarathjith MC, Diallo C, Berger A, Calatayud C, De Bellis F, Rami J, Vaksman M, Segura V, Pot D and De Verdal H, 2025. Factors influencing phenomic prediction: A case study on a large sorghum back cross nested association mapping population. The Plant Phenome Journal, 8, e70051. doi: <http://dx.doi.org/10.1002/ppj2.70051>
- Braun V, Zhu X, Meyenberg C, Hahn V, Maurer HP, Würschum T and Thorwarth P, 2025. Phenomic prediction can be improved by optimization of NIRS preprocessing. Plant Breeding, 144, 454-469. doi: <http://dx.doi.org/10.1111/pbr.13265>

- Chen T and Guestrin C**, 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 785-794. doi: <http://dx.doi.org/10.1145/2939672.2939785>
- Caporaso N, Whitworth MB and Fisk ID**, 2018. Near-infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews*, 53, 667-687. doi: <http://dx.doi.org/10.1080/05704928.2018.1425214>
- Fox GP, Onley-Watson K and Osman A**, 2002. Multiple linear regression calibrations for barley and malt protein based on the spectra of hordein. *Journal of the Institute of Brewing*, 108, 155-159. doi: <http://dx.doi.org/10.1002/j.2050-0416.2002.tb00534.x>
- Geyer M, Mohler V and Hartl L**, 2022. Genetics of the inverse relationship between grain yield and grain protein content in common wheat. *Plants*, 11, 2146. doi: <http://dx.doi.org/10.3390/plants11162146>
- Gous PW, Martin A, Lawson W, Kelly A, Fox GP and Sutherland MW**, 2012. Correlation between NIRS generated and chemically measured feed quality data in barley (*Hordeum vulgare*) and potential use in QTL analysis identification. *Euphytica*, 188, 325-332. doi: <http://dx.doi.org/10.1007/s10681-012-0644-x>
- Koekemoer FP, Labuschagne MT and Van Deventer CS**, 1999. A selection strategy for combining high grain yield and high protein content in South African wheat cultivars. *Cereal Research Communications*, 27, 107-114. doi: <https://doi.org/10.1007/BF03543926>
- Monaghan JM, Snape JW, Chojewski AJS and Kettlewell PS**, 2001. The use of grain protein deviation for identifying wheat cultivars with high grain protein concentration and yield. *Euphytica*, 122, 309-317. doi: <http://dx.doi.org/10.1023/A:1012961703208>
- Mosleth EF, Wan Y, Lysenko A, Chope GA, Penson SP, Shewry PR and Hawkesford MJ**, 2015. A novel approach to identify genes that determine grain protein deviation in cereals. *Plant Biotechnology Journal*, 13, 625-635. doi: <http://dx.doi.org/10.1111/pbi.12285>
- Nigro D, Gadaleta A, Mangini G, Colasuonno P, Marcotuli I, Giancaspro A, Giove SL, Simeone R and Blanco A**, 2019. Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. *Planta*, 249, 1157-1175. doi: <http://dx.doi.org/10.1007/s00425-018-03075-1>
- Oury FX and Godin C**, 2007. Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? *Euphytica*, 157, 45-57. doi: <http://dx.doi.org/10.1007/s10681-007-9361-1>
- Paina C and Gregersen PL**, 2023. Recent advances in the genetics underlying wheat grain protein content and grain protein deviation in hexaploid wheat. *Plant Biology*, 25, 661-670. doi: <http://dx.doi.org/10.1111/plb.13534>
- Polisetty R and Chandra R**, 1989. Grain protein accumulation during grain development in barley (*Hordeum vulgare* L.). *Journal of Agronomy and Crop Science*, 163, 57-61. doi: <http://dx.doi.org/10.1111/j.1439-037X.1989.tb00742.x>
- Rapp M, Lein V, Lacoudre F, Lafferty J, Müller E, Vida G, Bozhanova V, Ibraliu A, Thorwarth P, Piepho HP, Leiser WL, Würschum T and Longin CFH**, 2018. Simultaneous improvement of grain yield and protein content in durum wheat by different phenotypic indices and genomic selection. *Theoretical and Applied Genetics*, 131, 1315-1329. doi: <http://dx.doi.org/10.1007/s00122-018-3080-z>
- Richard R, Lovegrove A, Tosi P, Casebow R, Poole M, Wingen LU and Shewry PR**, 2025. Genetic analysis of grain protein content and deviation in wheat. *Journal of Cereal Science*, 121, 104099. doi: <http://dx.doi.org/10.1016/j.jcs.2024.104099>
- Robert P, Auzanneau J, Goudemand E, Oury FX, Rolland B, Heumez E, Bouchet S, Le Gouis J and Rincet R**, 2022. Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theoretical and Applied Genetics*, 135, 895-914. doi: <http://dx.doi.org/10.1007/s00122-021-03998-6>
- Savitzky A and Golay MJ**, 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36, 1627-1639. <http://dx.doi.org/10.1021/ac60214a047>
- Setotaw TA, Wilde P, Bernal Vasquez AM, Gebretsadik R, Lakew B, Shiferaw T, Difabachew YF and Jannink JL**, 2024. The potential of phenomic selection for the joint Ethiopian malt barley breeding program. *The Plant Phenome Journal*, 7, e70003. doi: <http://dx.doi.org/10.1002/ppj2.70003>
- Simmonds NW**, 1995. The relation between yield and protein in cereal grain. *Journal of the Science of Food and Agriculture*, 67, 309-315. doi: <http://dx.doi.org/10.1002/jsfa.2740670306>
- Thorwarth P, Piepho HP, Zhao Y, Ebmeyer E, Schacht J, Schachschneider R, Kazman E, Reif JC, Würschum T and Longin CFH**, 2018. Higher grain yield and higher grain protein deviation underline the potential of hybrid wheat for a sustainable agriculture. *Plant Breeding*, 137, 326-337. doi: <http://dx.doi.org/10.1111/pbr.12588>