The role of wheat flour minor components in predicting water absorption

Laura Rezette, Kamal Kansou, G. Della Valle, Sophie Le Gall, Luc Saulnier

The role of wheat flour minor components in predicting water absorption

Laura Rezette, Kamal Kansou, Guy Della Valle, Sophie Le Gall, Luc Saulnier

PII: S0308-8146(24)02882-6

DOI: https://doi.org/10.1016/j.foodchem.2024.141232

Reference: FOCH 141232

To appear in: *Food Chemistry*

Received date: 17 April 2024

Revised date: 14 August 2024

Accepted date: 9 September 2024

Please cite this article as: L. Rezette, K. Kansou, G.D. Valle, et al., The role of wheat flour minor components in predicting water absorption, *Food Chemistry* (2024), https://doi.org/10.1016/j.foodchem.2024.141232

The role of wheat flour minor components in predicting water absorption

Laura Rezette[1], Kamal Kansou[1], Guy Della Valle[1], Sophie Le Gall[1,2], Luc Saulnier[1]

[1]INRAE, UR 1268, Biopolymers, Interactions & Assemblies (BIA), F-44316 Nantes, France
[2]INRAE, PROBE Research Infrastructure, BIBS Facility, F-44316 Nantes, France

Abstract

Water absorption in wheat flour is a crucial parameter for optimizing bread-making processes. The determinants of wheat flour water absorption were investigated through the analysis of 28 compositional and technological properties of 150 wheats grown in France. A multiple linear regression approach was used to predict the water absorption, selecting the best model through successive examination of Bayesian Information Criterion, Variance Inflation Factor and minimizing the total number of variables.

A model with protein content, soluble starch, damaged starch and specific viscosity from water extractable arabinoxylans was identified as the best trade-off between the number of variables and the predictive performances among all possible models. Soluble Starch, varying between 1.11 and 6.21 g/100g flour a new criterion measured alongside water-extractable arabinoxylans content, varying between 0.26 and 0.86 g/100g flour, shows significant potential to predict water absorption compared to damaged starch.

Keywords

Wheat, water absorption, bread-making, lipids, arabinoxylans, damaged starch, multi-factorial analysis.

Abbreviations:

A: Arabinose

AGP: Arabinogalactan Peptide

AIM: Alcohol Insoluble Material

AX: Arabinoxylans

A.X.TOT: Arabinose on Xylose ratio from Total Arabinoxylans

AX-TOT: Total Arabinoxylans content

A.X.WE: Arabinose on Xylose ratio from Water-Extractable Arabinoxylans

A.X.WU: Arabinose on Xylose ratio from Water-Unextractable Arabinoxylans

C16.TOT: Total palmitic acid C16 content

C18.TOT: Total stearic acid C18 content

C181n7.TOT: Total vaccenic acid C18:1n-7 content

C181n9.TOT: Total oleic acid C18:1n-9 content

C182n6.TOT: Total linoleic acid C18:2n-6 content

C183n3.TOT: Total alpha-linolenic acid C183n-3 content

C16.NS: Non-Starch palmitic acid C16 content

C18.NS: Non-Starch stearic acid C18 content

C181n7.NS: Non-Starch vaccenic acid C18:1n-7 content

C181n9.NS: Non-Starch oleic acid C18:1n-9 content

C182n6.NS: Non-Starch linoleic acid C18:2n-6 content

C183n3.NS: Non-Starch alpha-linolenic acid C183n-3 content

CV: Coefficient of Correlation

D.Gluten: Dry Gluten

EI: Elasticity Index

FA-TOT: Total Fatty Acid

Gli: Gliadins

GluI: Insoluble Glutenins

GluS: Soluble Glutenins

GluT: Total Glutenins

HFN: Hagberg Falling Number

IV.AX: Intrinsic Viscosity of Water-Extractable Arabinoxylans

Prot: Protein content in wheat flour

$Prot_G$: Protein content in wheat grain

SS: Soluble Starch content

SD: Damaged Starch measured with iodine absorption

SV.AX: Specific Viscosity of Water-Extractable Arabinoxylan

TWG: Thousand Weight Grain

UPP: Unextractable Polymeric Protein

W: Baking Strength

WA: Water Absorption

WE-AX: Water-Extractable Arabinoxylan content

W.Gluten: Wet Gluten

WU-AX: Water-Unextractable Arabinoxylans

X: Xylose

1. Introduction

The distribution of water within the flour components is important due to its impact on gluten network development, which primarily governs dough properties. Even small changes in water content can affect significantly the dough behaviour, underscoring the importance of precise control over water addition during mixing. Water absorption (WA) is a key parameter for assessing wheat flour and is commonly determined using farinograph or similar devices to achieve a mixing consistency considered optimal.

WA can vary from 53.9% to 65.5 % (Sapirstein et al., 2018) and both proteins and damaged starch significantly impact it. According to (Greer & Stewart, 1959), WA is positively correlated with protein and damaged starch contents, with these two variables collectively accounting for 90% of its variation. Notably, damaged starch exerts a stronger influence on water absorption, a finding supported by other authors (Dodds, 1971; Tara et al., 1972; Tipples et al., 1978). Most studies also found a significant correlation between protein levels and damaged starch. However, Tara et al., (1972) and Dexter et al., (1994) reported that WA is only slightly affected by protein content. These studies are limited to total protein content, making no distinction between gliadin and glutenin which built the gluten network. Each of these fractions has different properties and impacts on the gluten-network (Barak et al., 2013; Dhaka & Khatkar, 2015; Park et al., 2006;), which may affect flour water absorption.

WA is also influenced by other components as recently confirmed by Sapirstein et al. (2018) who found that water-extractable arabinoxylans (WE-AX) positively impact WA. This finding aligns with insights from Andersson et al. (1993), who first emphasized that the content and the composition of arabinoxylans (AX) play a crucial role in WA. AX are known for their influence in bread-making (Courtin & Delcour, 2002; Marion & Saulnier, 2020;; Zhang et al., 2019). More precisely, the water-

unextractable fraction (WU-AX) has a high-water retention capacity (Jelaca & Hlynka, 1971; Meuser & Suckow, 1986), while the WE-AX fraction is recognized for its high viscosifying properties (Saulnier, 2019). However, AX are credited with various effects on water absorption and bread-making in general, yet the precise impact of the variability in their amount in wheat flour remains unclear. Most studies on the topic artificially introduce these components into the flour at high concentrations (Biliaderis et al., 1995; Zhang et al., 2019) or modify them with the addition of xylanases (Courtin & Delcour, 2002; Rouau et al., 1994) during the process. Studying their natural impact on WA without modification proves challenging due to the low concentration of AX in flours (1.3 to 2.7 % (Saulnier, 2019), compared to the major components.

In addition to AX, another minor component of flour, lipids, ranging from 2 to 2.5% (Pareyt et al., 2011), have been shown to influence bread-making. In particular, lipids interact with proteins, such as puroindolines (Marion et al., 2003) and the gluten network where they alter disulphide bridges (Nishiyama & Kuninori, 1987). They also interact with starch, forming complexes with amylose (Morrison, 1988). Lipids could have an indirect effect on water absorption through their interactions with proteins and starch.

Our hypothesis is that flour minor components explain a significant fraction of the natural variability of the flour water absorption, even though proteins and starch are the main contributors. The main objective of this work is to improve WA prediction and to ascertain the role of each flour component, especially how the natural variability of minor components can affect WA, which is important parameter for the control of the breadmaking process. To this end, 150 wheat samples (harvested in 2020 and 2021) were characterized for their content in minor components (lipids and AX), as well as the detailed composition of proteins, in addition to protein and damaged starch contents, and classical technological parameters. Linear regressions were performed in combination with a variable selection methodology based on the Bayesian Information Criterium score (BIC), to identify the most significant variables to predict WA.

This work should help identifying whether assessment of lipids and AX, can provide valuable insights for controlling wheat flour properties.

2. Materials and methods

2.1. PLANT MATERIAL

A total of 150 wheat samples harvested in 2020 and 2021, originating from 67 different varieties and 26 growing locations across France with diverse agro-pedo-climatic conditions were used to cover a wide range of end-use quality. Among the 67 wheat varieties, 11 are classified for biscuit applications, 32 as 'improver wheat' with a high protein content, and the remaining varieties, which comprise the majority of samples, are classified for bread production. These samples were provided by Arvalis (Boigneville, France), Limagrain (Riom, France), and Axiane Meunerie (Val d'Arnast, France). The grains were milled into white flour (0.55% of ashes in average, corresponding to Euro 550 flour type and American all-purpose flour, containing almost exclusively grain endosperm) with an experimental mill (MCKA, Bülher, Switzerland) in batches of 10 kg.

The flours were kept for 20 days at room temperature after milling and then were frozen at -20°C until use.

The samples were unfrozen over night at room temperature prior to analysis.

2.2. FLOUR AND DOUGH CLASSICAL CHARACTERIZATION

The usual analyses of grain, including protein content ($Prot_G$) and Hardness, were carried out using near-infrared spectroscopy (NIRS). The protein content was determined by NIRS according to NF EN 15948 and the NIRS calibration for Hardness was conducted using the Particle Size Index (PSI). The Hagberg Falling Number (HFN) was determined using Falling Number PERTEN FN 1500 equipment according to ISO 3093:2009. Damaged starch measurements (SD) was measured with the SDmatic (SDmatic, Chopin Technologies, Villeneuve la Garenne, France) using the amperometric method

according to NF EN ISO 17715:2015, with results expressed in Chopin Units corrected based on protein and moisture content (UCDc). The Glutomatic® system (PerkinElmer, Waltham, USA) was used to determine Wet Gluten (W.Gluten), Dry Gluten (D.Gluten) and Gluten Index according to NF EN ISO 21415-2). The Alveograph® (Chopin Technologies, Villeneuve la Garenne, France) was used to determine dough characteristics: tenacity (P), extensibility (L), tenacity to extensibility ratio (P/L), baking strength (W), and elasticity index (IE), according to NF EN ISO 27971. Mixolab (Chopin Technologies, Villeneuve la Garenne, France) was used to determine water absorption (WA), Development Time, Stability and Weakening degree according to NF V 03-765. The Mixolab equipment allows for obtaining values comparable with existing Farinograph® equipment, with a much smaller sample size. WA is defined as the percentage of water required for the dough to produce a torque of 1.1 Nm.

2.3. ADVANCED CHARACTERIZATION OF FLOUR COMPOSITION

2.3.1. Polysaccharides analysis

Alcohol Insoluble Materials extraction

Flour samples were prepared as Alcohol Insoluble Materials (AIMs) using an automated extraction method with an accelerated solvent extraction unit ASE® 350 (THERMO, CA, USA). Flour samples (2 g) were extracted using 80% ethanol at a flow rate of 2 mL/min in 22 mL cells of ASE® 350. The conditions for the ASE extraction were set at 100°C with a flow time of 20 min, followed by a rinse volume of 150%, and a purge time (N_2) of 30 seconds. AIMs were recovered and dried at 40 °C for three hours and then overnight under vacuum over P_2O_5 before grinding with a knife grinder (IKA Tube Mill 100 control, IKA-werke GmbH & Co. KG, Staufen, Germany) for 30 seconds and weighing.

Aqueous extracts

Each AIM was weighted (1g) in 15 mL falcon tubes. Ultrapure water (4 mL) was added and the contents of the tube were thoroughly mixed using a vortex mixer. The tubes were then shaken

overnight (16h) using a circular shaker (Multi reax, Heidolph; at 1700 rpm) placed in an oven at 40°C. The tubes were then mixed again using a vortex and centrifuged (2200 g, 30 minutes, 25°C). The supernatant (2 mL) was immediately transferred to 2 mL Eppendorf.

Determination of polysaccharides content

Constitutive monosaccharides were analysed by gas chromatography (GC) based on (Englyst & Cummings, 1988) method after AIM acid hydrolysis.

For the total carbohydrate content, 5 mg of AIM were weighed and hydrolysed with 1M H_2SO_4 at 100°C for 2h. For the water-soluble carbohydrate content, AIM aqueous extracts were hydrolysed in the same conditions using 0.2mL of centrifuged supernatant. Inositol was used as internal standard.

Released monosaccharides were then converted into their alditol acetate equivalents as previously described (Hoebler et al., 1989) and analysed by GLC on a TG-225MS column (Trace GC Ultra, THERMO; temperature 205 °C, carrier gas H_2). Each sample was analysed in triplicate and the total arabinoxylan content (TOT-AX) and water-extractable arabinoxylan content (WE-AX) was calculated as the sum of arabinose and xylose. The arabinose content was corrected for the presence of arabinogalactan peptide (AGP) on the basis of an arabinose to galactose ratio of 0.7 and with the assumption that all of the arabinose of AGP is present in the aqueous extract (Fincher & Stone, 1974; Loosveld et al., 1998). The water-unextractable arabinoxylan content (WU-AX) was calculated by subtracting WE-AX from TOT-AX. The water-extractable glucose or soluble starch (SS) was measured from the AIM aqueous extracts. Note that the small fraction of water-extractable glucose originating from mixed-linked beta-glucan is considered negligible. Arabinose / Xylose ratio (A/X) was also calculated for total, water-unextractable and water-extractable arabinoxylans (respectively A.X.TOT, A.X.WU and A.X.WE). A/X indicates the degree of substitution of the xylan backbone with arabinose residues.

Determination of polysaccharides physicochemical characteristics

AIM aqueous extracts (1.8 mL) were treated with 3 units of α-amylase thermostable (Bacillus sp., product code E-BSTAA, Megazyme) overnight at 30°C. Then, aqueous extract were filtered over 0.45 µm membrane and injected on a high-performance size exclusion chromatography (HPSEC) system (OMNISEC RESOLVE-REVEAL - Malvern Panalytical- Malvern, UK) using a Viscotek AGuard precolumn (50 x 6 mm) and a Viscotek A4000 column (300 x 8 mm – Malvern Panalytical- Malvern, UK) maintained at 35°C and eluted with 50 mM sodium nitrate at a flow rate of 0.7 mL/min. Measurements were performed using a differential refractometer (OMNISEC REVEAL), a multi angle laser light scattering detector (λ = 660 nm, 44°, 60°, 76°, 90°, 108°, 124°, 140°, VISCOTEK SEC-MALLS 9) and a differential pressure viscometer (OMNISEC REVEAL). Detectors responses were calibrated with a pullulan standard having narrow molecular mass distribution (weight-average molar mass = 40,611 Da, number-average molar mass = 38,931 Da, IV = 23,6 mL/g at 30°C in 0.1 M sodium nitrate, refractive index increment dn/dc = 0.147 mL/g). Data analyses were carried out using OmniSec version 11.32 software (Malvern Panalytical) and a dn/dc value of 0.146 mL/g was used for WE-AX (Dervilly et al., 2000).

The peak eluted within the range of 5.4 – 8.1 mL was integrated to calculate the concentration of water-extractable arabinoxylans ($WEAX_{HPSEC}$) from the refractive index signal. $WEAX_{HPSEC}$ is highly correlated with WE-AX concentration determined by chemical analysis (r=0.91). SV.AX, the Differential Pressure (DP) measured across the capillary bridge of the viscometer was integrated within the elution range of 5.4-8.1. SV.AX is related to the specific viscosity η_{sp} of the AIM aqueous extract, intrinsic viscosity of WE-AX (IV.AX) and $WEAX_{HPSEC}$ according to the following equations (Haney, 1985):

$$\eta_{sp} = \frac{4\mathrm{SV.AX}}{\mathrm{IP} - 2\mathrm{SV.AX}} = \mathrm{IV.AX} * \mathrm{WEAX_{HPSEC}}$$

Where IP is the inlet pressure of the viscosimeter. Since IP is almost constant and far higher than SV.AX, $\eta_{sp} \propto$ SV.AX. Hereafter, SV.AX is considered as the specific viscosity associated with arabinoxylans. Due to the large number of samples and the time required for analysis, single analyses

were performed. To ensure repeatability and reliability, reference flour was repeatedly analyzed, confirming that values for SV.AX and IV.AX had a coefficient of variation (CV) of less than 5%.

2.3.2. Lipids analysis

Extraction of non-starch fatty acids

Lipid were isolated with a semi-automated extraction method using accelerated solvent extraction unit ASE® 350 (THERMO, CA, USA). Flour samples (500 mg) were treated using hexane/propanol solvent (3:2, v/v) at a flow rate of 2 mL/min in 22 mL cells of ASE 350. The conditions for the ASE extraction were set at room temperature with a flow time of 3 min. The rinse volume at the end of the extraction was 100% of cell volume, and the purge time (N_2) was set to 30 seconds.

The non-starch (NS) fatty acids were recovered in the 50 mL hexane/propanol fraction after ASE extraction.

Subsequently, 5 mL of the fraction was transferred into 8 mL screw cap tube and dried under reduced pressure at 45°C, using a Genevac (SP Scientific, Warminster, PA, USA). Three fractions in each extract were recovered to enable triplicate analyses.

Determination of fatty acid content

The non-starch (NS) and total (TOT) fatty acids content of each flour was determined by gas chromatography following transmethylation of lipids as described by Welch (1977). An amount of 10-12 mg of flour samples was weighted into 8 mL screw cap tube. 2,2-dimethoxypropane (100 µL) was added and incubated for 30 min to chase water. Internal standard (margaric acid, 1 mg/mL) was then added and mixed with 2% sulphuric acid in methanol (4 mL). The tubes were sealed and heated for two hours using a Swing XL workstation (Chemspeed® technologies AG) to make 3 shaking every 20 minutes during the first hour. A similar procedure was applied on dried hexane/propanol extracts but without the 2,2-dimethoxypropane step.

Ultrapure water (2 mL) and cyclohexane (1.5 mL) were added and each tube was mixed for 10 seconds using a vortex mixer. After cooling overnight at 4°C, 1 mL of the upper layer was withdrawn, transferred into vials and then 1 µL was analysed by GLC (Clarus 690 GC, PERKINELMER, temperature 250°C, carrier gaz H_2) on a DB225 J&W Scientific column, for the separation and measurement of the fatty acid methyl ester peak.

The amount of fatty acids bound to starch was calculated by subtracting the amount of non-starch fatty acids as detected in the hexane/propanol extracts from the total fatty acid content measured in the flour.

2.3.3. Proteins analysis

Molecular size distribution of gluten proteins was studied by high-performance size exclusion chromatography (HPLC Alliance, Waters) as described by (Morel et al., 2000) by performing two sequential extractions. Flour (160 mg) was placed in a 50 mL Nalgene centrifuge tube with 20 mL of extracting buffer solution (0.1 M sodium phosphate buffer solution (pH 6.9), 1% sodium dodecyl sulphate). Tube was shaken 80 min at 60°C (Heidolph Reax 2, setting 5) before centrifugation at 25°C, 18,000 rpm (Beckman, JA 20, fixed-angle rotor). The supernatant was collected and 2 mL stored in HPLC vial at -18°C. The remaining flour pellet was dispersed in 5 mL of extracting buffer and sonicated at ambient temperature during 180 s at 30% of the nominal power (50 W, 20 Hz) (VibraCell 72434; Bioblock, Illkirch, France). The tube was centrifuged as above and 2 mL of this second extract was stored at -18°C. The two supernatants were injected (20 µL) onto a size exclusion column TSKgel® G4000- SW_{XL} (7.5 mm × 30 cm, Tosoh) coupled to a TSKgel G2000SW_{xl}-G4000SW_{xl} guard column (6 mm, 4 cm). Elution was performed at 0.7 mLl/min with 0.1 M sodium phosphate buffer solution (pH 6.9), 0.1% SDS. The column was calibrated with seven protein standards (PSS-PROKIT, Agilent). The UV signal at 214 nm was recorded and expressed in g/L considering a specific extinction coefficient of 18.51 L/g/cm for the wheat protein. The chromatogram of the first extract was integrated considering five fractions of increasing elution times (F1 to F5). Fractions F1 and F2, assigned to the SDS-soluble

glutenin polymers, comprised elution times from the void volume to 630,000 g/mol and from then to 116,000 g/mol. Protein species in the range of 116,000 to 65,000 (F3) and 65,000 to 21,000 g/mol (F4) were assigned to omega and then to gamma, beta and alpha gliadins. The last eluting fraction (F5, Mw < 21,000 g/mol) gathered the flour water-soluble proteins. The SDS-insoluble glutenin polymers fraction (Fi) was obtained from the total area of the second extract chromatogram. The total, soluble and insoluble glutenin polymers (referred to as GluT, GluS, and GluI respectively) contents in flour, along with the gliadins content (Gli), and unextractable polymeric protein (UPP) were determined as outlined by Baudouin et al. (2020). Due to the large number of samples and the time required for analysis, single analyses were performed. To ensure repeatability and reliability, reference flour was repeatedly analyzed, confirming that measurements had a coefficient of variation (CV) of less than 5%.

2.4. DATA TREATMENT FOR WA PREDICTION

2.4.1. Variable pre-selection

The raw dataset with composition and technological data for 150 wheat samples, was reduced to 144 due to missing data and obviously aberrant measurements for six wheat samples.

A wide range of flour composition analyses was performed on the wheat samples, of which twenty-four measurements were selected to compose the working dataset. The selection process basically removed the factors highly correlated with another. The detailed selection process is available in supplementary data (S1).

Finally, the composition variables considered for the study were:

- For proteins: Protein content (Prot), Total Glutenin content (GluT), Soluble Glutenin content (GluS), Unextractable Polymeric Protein (UPP), Gliadin to total Glutenin ratio (Gli.GluT) and Gliadin to soluble Glutenin ratio (Gli.GluS);

- For lipids: C16.TOT, C18.TOT, C181n7.TOT, C181n9.TOT, C182n6.TOT, C183n3.TOT representing the total (TOT) content of palmitic acid, stearic acid, vaccenic acid, oleic acid, linoleic acid and alpha-linolenic acid respectively, and C16.NS, C181n7.NS, C182n6.NS, C183n3.NS representing the Non-Starch (NS) content of palmitic acid, vaccenic acid, linoleic acid and alpha-linolenic acid respectively.
- For non-starch polysaccharides: Total arabinoxylan content (TOT-AX), Water-Extractable arabinoxylan content (WE-AX) and their respective Arabinose / Xylose ratio (A.X.TOT and A.X.WE), Intrinsic Viscosity of Water-Extractable Arabinoxylans (IV.AX) and Specific Viscosity of Water-Extractable Arabinoxylans (SV.AX)
- For starch: damaged starch measured by iodine absorption (SD) and Soluble Starch (SS)

2.4.2. Model selection

Statistical analyses were performed with R v4.1.3. The model selection was carried out with the help of the *regsubset* function of the *leaps* R package. The principle consists in selecting the best multiple linear regression models of the form $Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\varepsilon$ where Y represents the flour Water Absorption (WA) and X_1, X_2, …, X_n the combination of standardized wheat variables and their associated coefficients β_1, β_2, …, β_n, and ε the error.

The best model was selected through a successive examination of three criteria (Fig. 1):

1. Minimizing the BIC (Bayesian Information Criterion) score, which serves as a model general performance indicator accounting for the number of variables;
2. A VIF (Variance Inflation Factor) value below five, ensuring a low multicollinearity level among the variables, in agreement with the general usage;
3. Minimizing the total number of variables, when the two first criteria are met. The goal is to avoid over-learning due to an excess of variables.

The BIC allows ranking the possible regression models according to maximum likelihood principle, while penalising models with too many variables. The BIC score is defined as:

For each model (i) separately, the BIC score is defined as:

$$BIC_i = -2\,Log\,L_i + p_i \log n$$

Where L_i and p_i are the likelihood and the number of parameters for each model, and n is the number of observations. A smaller value indicates a preferable model.

The models were checked for multi-collinearity using the VIF function of *car* R package. In addition, the R^2 is provided to indicate the proportion of variance explained by the regression model.

Finally, the predictive performance was assessed with a repeated K-fold cross validation. This method involves randomly partitioning the samples into k groups (folds) of equal size (when possible). (k-1) groups are dedicated for training the regression model, while the remaining group is used for validating. This procedure is repeated k times, with each iteration using a different group for validation. The *trainControl* function using the *repeatedcv* method from the *caret* library in R was employed to conduct 10-fold cross validation with 3 repetitions to explore alternative divisions into 10 folds.

3. Results and discussion

3.1. SAMPLE VARIABILITY

Descriptive statistics of the dataset are provided in Table. 1. The rather wide range of values and the variability indicate that the wheat displays various technological behaviors and compositions, which are sought-after characteristics for this study.

In terms of composition, the flour protein content Prot (6.1 – 14.0 g/100 g) and grain protein content ProtG (8.9 – 17.6 g/100 g) are highly correlated (r=0.98), which is commonly observed. The range of ProtG is similar to Dexter et al.(1994) measurements with 10.8 – 18.5 g /100 g for approximately 200

individual farmers' deliveries but is wider than those reported by Dowell et al. (2008) (11.4 –15.8 g/100 g; 148 samples with a broad spectrum of bread quality) or by Sapirstein et al. (2018) (11.1 – 15.1 g/100 g; 78 samples comprising 59 different genotypes). The range of values for Gliadin to total Glutenin ratio Gli.GluT (0.86 – 1.46) is wider than in Dhaka & Khatkar, (2015) who reported values from 0.75 to 1.16 for 15 varieties presenting distinct performances for bread-making.

The range of values for damaged starch measured by iodine absorption SD (8.7 – 23.9 UCDc) are in line with Dragan et al. (2012) and Golea et al. (2023). They reported, respectively, SD values from 19.55 to 29.05 UCDc for 18 different wheat varieties and from 1.40 to 14.60 UCDc for 66 wheat samples collected in different regions.

The total arabinoxylan content AX-TOT (1.27 – 2.69 g/100 g), as well as the water-extractable fraction content WE-AX (0.26 – 0.86 g/100 g), are consistent with the results obtained by Gebruers et al., (2008) (AX-TOT: 1.35 – 2.75, WE-AX: 0.15 – 1.40; 176 different varieties) and Sapirstein et al.(2018) (1.08 – 1.80 g/100 g for AX-TOT and 0.23 – 0.46 g/100 g for WE-AX). Moreover, mean values for AX-TOT (1.83 g/100 g) and WE-AX (0.5 g/100 g) are close to those reported by Selga et al., (2023) with 197 wheat samples.

The proportion of each fatty acids is in line with previous works (De La Roche et al., 1975; Prabhasankar & Haridas Rao, 1999). Palmitic acid (C16.TOT) and linoleic acid (C182n6.TOT) are the most abundant fatty acids in the wheat flour but C16.TOT has one of the lowest Coefficient of Variation (CV) among fatty acids.

In terms of technological characteristics, range of values for W.Gluten (10.5 – 40.6 g/100g) and Dry Gluten (3.5 – 13.5 g/100g) was also wider than in Dhaka & Khatkar, (2015) who reported the following intervals: W.Gluten: 21.5 – 35.5 g/100g, D.Gluten: 8.4 – 12.4 g/100g. The range of values for Hardness corresponds to what is found in the literature, by encompassing soft and hard grains (Rakszegi et al., 2010). The range of values for the Hagberg Falling Number HFN is rather large (190 - 462 s), although narrower at the lower limit when compared to the results of Mangan et al.(2016) who observed

values below 190 s with a dataset of about the same size. These technological variables are included in the analysis as they take into account parameters that may have an impact on Water Absorption (WA), such as grain characteristics and particle size for Hardness, amylase activity for HFN, and the quality of the gluten network for W.Gluten and D.Gluten. Moroever, the lower bound of the range of values for WA (50.4 – 63%) is rather noteworthy, compared with results from the literature (53.9 – 65.5% in Sapirstein et al. (2018)).

Alveograph parameters are excluded from the analysis as they are determined with a fixed amount of water added to the dough (50%). However, they are given as complementary technological variables to illustrate the variability of the dataset's breadmaking potential (see Table 1). Indeed, the ranges of values for P/L (0.13 -2.54), W (68.10^{-4}J – 546.10^{-4}J) and Ie (32.8 – 73%) are equivalent to, or even broader than, those reported by Jødal & Larsen (2021) (P/L: 0.15 – 2.28, W: 62.10^{-4}J – 352.10^{-4}J, Ie: 31.1 – 61.5) with 532 individual pressure curves.

These results highlight the high variability of the wheat samples in term of composition and technological properties, and therefore the richness of the working dataset for investigating the potential of wheat for breadmaking.

3.2. WHEAT COMPOSITION DETERMINANTS OF THE WATER ABSORPTION

3.2.1. Exploring correlations in the dataset

At first, a heatmap illustrating the correlations between all variables in the dataset was generated (Fig. S2). The strongest correlations of WA with composition variables are found for Proteins (Prot), Damaged Starch (SD) and Soluble Starch (SS); corresponding scatter plots are presented in Fig.1a, b and c, respectively.

The rather high correlations of WA with Prot and SD are expected, since both are commonly used variables to predict WA. Damaged starch typically absorbs 2 to 4 times its weight in water, compared to only 0.4 times for intact granular starch. It is the primary variable associated with WA in many studies (Dodds, 1971; Greer & Stewart, 1959; Tara et al., 1972; Tipples et al., 1978). No direct

correlation was found between Prot and damaged starch (SD) (r=-0.07), in agreement with results previously reported by Tara et al. (1972).

Proteins (Prot) exhibits a stronger correlation with WA (r=0.63, Fig.1a) than SD (r=0.56, Fig.1b), which appears in contradiction with most previous results (Dodds, 1971; Greer & Stewart, 1959; Tara et al., 1972; Tipples et al., 1978). Surprisingly, soluble starch (SS) presents the highest correlation with WA (r=0.72, Fig.1c). SS is actually a measurement performed in our laboratory, and would correspond to the glucose originating from starch solubilized during the overnight water extraction of flour AIM (unpublished results). Interestingly, SS correlates much better with hardness (r=0.76, Fig.1e) than SD does (r=0.54, Fig.1f), despite the significant influence of hardness on damaged starch formation (Rakszegi et al., 2010). These results suggest that SS could serve as a reliable marker for damaged starch. SS and SD are correlated (r=0.7; Fig.2d) but the correlation level was lower than expected, meaning that they likely measure different characteristics of the damaged starch. The heatmap (Fig. S2) complemented with hierarchical clustering, actually groups Hardness, WA and SS, indicating a strong connection between them.

In addition, to protein and damaged starch contents, correlations were observed between WA and other composition variables (see Fig. S2). For example, C182n6.TOT (r=0.58), and C183n3.TOT (r=0.58), exhibited slightly better correlations with WA than SD. These two fatty acids are mainly coming from polar lipids associated to starch.

3.2.2. Modelling the influence of damaged starch and of protein content on water absorption

The effect of wheat components on WA was analyzed through multiple linear regression models. Since the dataset contains measurements with different units and scales, the values were first standardized so that to compare the relative significance of each influence. The resulting coefficients assigned to these variables, as well as the metrics values for model prediction, are reported in table 2.

The simple linear regressions between WA and composition variables SD, SS and Prot are **the primary models**, noted M_{SD}, M_{SS} and M_{Prot} respectively.

M_0, $M_{0.1}$, $M_{0.2}$, and $M_{0.3}$, named **usual models**, are multiple linear regression models combining Prot, SD and SS. M_0 which includes SD and Prot plays the role of benchmark, since Prot and SD are commonly used to predict WA. $M_{0.2}$ is a version of this model, combining Prot and SS, assuming that SS is also representative of damaged starch. $M_{0.1}$ combines the two variables related to damaged starch and finally, $M_{0.3}$ brings the Prot, SS and SD together.

As expected, the four usual models exhibit much better performances than the primary models as shown by the R^2 and BIC values on Table 2. Amongst the usual models, M_0 and $M_{0.3}$, exhibit the best performances (R^2=0.76 and 0.78 respectively). $M_{0.1}$ has the worst predictive performance, it is also the only models without Prot, the primary driver of WA prediction, as its regression coefficient is the highest in the more complete model, $M_{0.3}$. With the exception of $M_{0.3}$, SS demonstrates the highest regression coefficients among the models in which it is included, but although SS provides also the most effective primary model, M_{SS}, it does not improve the model's performances when combined with Prot and/or SD. Indeed, M_0 (only usual model without SS) performs better than $M_{0.2}$, and $M_{0.3}$ (most complete usual model) fails to improve significantly the prediction compared to M_0.

3.2.3. Modelling the influence of composition variables on water absorption

In addition to the usual variables, Prot, SD and SS, the dataset contains measurements for protein-related variables (GluT, GluS, UPP, Gli.GluS, Gli.GluT, GluI), arabinoxylans-related variables (WE-AX, AXTOT, A.X.WE, A.X.TOT, IV.AX, SV.AX) and lipid-related variables (C16.TOT, C18.TOT, C181n7.TOT, C181n9.TOT, C182n6.TOT, C183n3.TOT, C16.NS, C181n7.NS, C182n6.NS, C183n3.NS), which are potential determinants for WA. A large number of regression model can be produced by combining all these variables. To select relevant models, the selection model method, based on the BIC score and described in section 2.4.2., has been applied and a graphical representation of the main results is presented in fig.2. The main guides for model selection are: 1) a small BIC value (negative index) which

indicates a better modelling performance, 2) a VIF score <5 to avoid strong multicollinearity amongst the variables, 3) a minimal number of variables when the two first criteria are met.

Following this method, a total of nine regression models has been selected, including M_{ss} and M_0. The other models, numbered from M_1 to M_7 are arranged in ascending order based on the number of explanatory variables, from three variables in M_1 to nine in M_7 (Fig.3). In Table 2, M_1 to M_7 are categorized as **composition models** and all their variables are significant (p-value<0.05). Additionally, an extra model, M_x, is provided to illustrate the best three-variable model that complements the reference model M_0. The recurring inclusion of variables across the selected models provides an indication of their relative importance as determinants for WA. Specifically, all composition models except M_x include Prot, SS and SV.AX as explanatory variables.

Subsequently, SD appears in all models except M_1, while the Arabinose to Xylose ratio from Water-Extractable Arabinoxylans (A.X.WE) is present in all the models except M_1 and M_2. The total steatic acid content (C18.TOT) is absent from M_1 to M_3, followed in order of importance by the Non-starch linoleic acid content (C182n6.NS), Unextractable Polymeric Protein (UPP) and the Intrinsic Viscosity of Water-Extractable Arabinoxylans (IV.AX).

Fig. 3 displays the BIC score against the number of variables for all the models. From one to three variables the BIC score decreases steeply, and above three variables, around a BIC score of -250, the decrease is moderate. The model M_2 achieve an interesting trade-off between the number of variables and the BIC score. With four variables, M_2 reaches a BIC score of -252 and a r of 0.92.

Surprisingly, the M_2 includes both SD and SS, which strengthens the idea that both measurements capture distinct aspects of the damaged starch and that both are determinants of WA. The contribution of SS is slightly higher than of SD in the composition models as indicated by the regression coefficients (Table 2). M_2 is also a model that includes an arabinoxylan variable, via SV.AX, besides starch and protein. SV.AX is also included in M_1, which is close to M_2 in Fig.3. M_1 performs slightly less well than M_2 (R^2=0.83 against R^2=0.85) and represents the best three-variable model with

a significant improvement compared to $M_{0.3}$ (R^2=0.78) which includes SD instead of SV.AX. Therefore, SV.AX is a minor component characteristic that has clearly improved the prediction of WA.

Fig. 4 reports the predictions of selected models computed through cross-validation. It is evident that M_2 significantly reduces the prediction error over the full range of WA compared to M_0, especially for low WA. The nine-variable composition model, M_7, improves the prediction of high WA compared to M_2. However, relative to the number of variables, it does not significantly improve the global prediction (R^2_{CV}=0.88 for M_7 compared to 0.84 for M_2).

3.2.4. Including technological properties of wheat in the model

The contributions to WA of the following technological measurements are examined in addition to the wheat composition:

- Hardness, usually strongly related to damaged starch
- Hagberg falling Number (HFN), often associated the amylase activity of the wheat grain
- Wet gluten, W.Gluten, and dry gluten, D.Gluten, which characterize the quality of the gluten network

The same model selection procedure has been applied. First, the BIC score is computed for generated models and results are reported in Fig. 5. It shows that HFN is not included in most models, suggesting that its contribution to WA is negligible.

Following the model selection explained in section 2.4.2, several models were discarded because of multicollinearity, i.e. VIF >5. The eight remaining models (highlighted on Fig. 5) include the models M_0, M_1, M_2 and M_3 and four new models, M_H, M_{H1}, M_{H2} and M_{H3} of respectively 1, 6, 8 and 9 explanatory variables. These four models all incorporate hardness, M_H being actually the simple linear regression between hardness and WA. With the exception of M_H, Prot has been systematically included in all models, resulting in the exclusion of D.Gluten and W.Gluten due to their VIF score exceeding five when Prot is involved. This reflects the high correlation amongst these three variables (Fig. S2). This result suggests that Prot exerts a more substantial influence on WA than D.Gluten or W.Gluten.

Furthermore, it suggests that WA is more dependent on the quantity of protein rather than its quality within the gluten network.

In agreement with the strong correlation between hardness and WA (r=0.77, S2), hardness comes out as a good explanatory variable for WA (Table 2). In M_{H1}, M_{H2} and M_{H3}, hardness is associated with SD and SS. The link between these three variables is further highlighted by the heatmap shown in Fig.S2). SS and SD remain important explanatory variables after Prot and before Hardness in M_{H1}, whereas SS is less significant than Hardness in M_{H3} (Table 2). The coefficients of SS in M_{H1}, M_{H2} and M_{H3} are lower than in the composition models, meaning that part of SS contribution to WA is now captured by Hardness. SD contribution in these models is not affected to the same extent.

Finally, except for M_H, incorporating Hardness in the models has not significantly improve the prediction of WA as shown in Fig. 3. Indeed, M_{H1} M_{H2} and M_{H3} perform only slightly better (BIC score = -289, and R^2=0.90 for the most complete model M_{H3}) than the composition models with the same number of variables, respectively M_4 M_6 and M_7. Prediction results displayed in Fig. 5.c.d. confirm that M_7 and M_{H3} perform similarly. To minimise the number of variables, the best regression model remains M_2, including Prot, SS, SD and SV.AX. This suggests that this combination of composition variables offers an effective solution to predict and explain WA, while the technological criteria investigated did not significantly enhance the accuracy of the model predictions.

3.2.5. Towards new predictor of WA

The water absorption of wheat flour is a complex and multifactorial phenomenon. Thus, any model relying on a single explanatory variable has exhibited low predictive performances. M_{ss} stands out as the most robust single-variable model with a r of 0.72 (Table 2), while the four-variable model M_2, a good compromise between the number of variables and the predictive performance, achieve a r of 0.92. M_2 includes the Protein content (Prot) and the damaged starch measured by iodine absorption method (SD), two usual measurements employed for WA prediction in the domain literature.

However, M_2 also included Soluble Starch (SS) and Specific Viscosity related to Water-Extractable Arabinoxylans (SV.AX), two less common variables for predicting WA.

SS, like SD, could serve as a marker of damaged starch. Amylose is likely to be more easily leached from a broken granule during the water extraction of flour than from a native starch (Wang et al., 2020). Consequently, the variation in the amount of water-soluble starch (SS) could reflect the proportion of broken granules, i.e., damaged starch, in the flour sample. Furthermore, SS exhibited a higher CV than SD, suggesting that this measurement is rather sensitive and potentially discriminant of the various wheat samples. Unlike SS, SD consistently exhibits VIF scores below 5 (as shown in Fig. 2 and Fig. 5), indicating its independence from the other variables. This supports the idea that these two measurements explain different aspects of damaged starch. However, considering that SS reflects released amylose, the average level of damaged starch in our study is 10-13% (assuming amylose represents 25-30% of the starch), which is higher than the average values typically reported, around 6-8% (Dodds, 1971). This discrepancy may be due to the Alcohol Insoluble Material procedure, which involves high-pressure ethanol treatment and additional grinding that could further damage the already broken granules, leading to increased solubilization of amylose and potentially amylopectin. Therefore, while SS is a useful indicator, it remains an indirect evaluation of damaged starch and may particularly overestimate higher values compared to standard measurements. Damaged starch is always measured indirectly, resulting in estimations that may vary between methods. Our findingsindicate that both measurements (SS and SD) reflect the starch state and, when used together, improve the prediction of flour water absorption.

Specific viscosity (SV.AX) is a physico-chemical property of the water phase controlled by the concentration of Water-Extractable Arabinoxylans (WE-AX) and their intrinsic viscosity (IV-AX), which is related to polymer size: SV.AX $\propto$ WE-AX × IV.AX. While AX concentration significantly influences the variation in specific viscosity, the variation in their molecular size also has an impact, as illustrated in Fig. 6. In this figure, the impact of molecular size of WE-AX (i.e., intrinsic viscosity; IV.AX) on specific

viscosity (SV.AX) is shown for two samples exhibiting the same amount of WE-AX. The inclusion of SV.AX in M_2 confirms the findings of Sapirstein et al.(2018) regarding the contribution of water extractable arabinoxylans to WA, but provides clarity regarding the significance of other feature apart from their concentration. Interestingly, it suggests that the influence stems from soluble arabinoxylans rather than their insoluble counterparts (included in TOT-AX) yet known for their high-water retention capacity (Marion & Saulnier, 2020). The relationship between the contents of water-extractable arabinoxylans (WE-AX) and total arabinoxylans (TOT-AX) in the starchy endosperm varies among cultivars. The proportion of WE-AX can range significantly, from 20% to 40% of the total AX content. Variation in WE-AX content in wheat flour, although primarily influenced by genotype, is also affected by the environment (Marion & Saulnier, 2020). In this respect, it is not known whether environmental factors influence to a greater extent the WE-AX content in flour or its molecular size, which both have an effect on the specific viscosity of flour water extract.

Regarding proteins, the total content Prot, appears to be sufficient for the prediction of WA, since further specific feature about the type of protein has not been selected in the models or has been discarded due to multicollinearity. Even if glutenin and gliadin fractions have different properties, their impact on WA is not distinguished. Moreover, the quality of the gluten network described by W.Gluten and D.Gluten variables does not seem to provide as much information as Prot in predicting WA.

Furthermore, the model selection process in this study has not pointed any critical fatty acids, suggesting a minimal role of the lipids for WA. Nevertheless, C18.TOT and C182n6.NS were included in about 80% of the models proposed by the BIC test (Fig. 2 and Fig. 5). Surprisingly, C18.TOT, did not present a VIF score above 5 in any model, suggesting independence from other fatty acids variables. On the contrary, although C182n6.TOT and C183n3.TOT, showed good simple linear correlations with WA, they were not the most commonly proposed fatty acids in the models. When included, they consistently had a VIF score above 5.

Finally, SS and SV.AX can be retained as new predictors of WA. It is particularly noteworthy that SV.AX, representing variation in a minority component of the flour – WE-AX – stands out significantly alongside the majority components in explaining WA.

4. Conclusion

The 150 wheat samples collected in this study exhibited a wide range of compositional and technological characteristics, facilitating the testing of various models to predict flour Water Absorption (WA). A model incorporating protein content (Prot), soluble starch (SS), damaged starch (SD) and specific viscosity associated with water extractable arabinoxylans (SV.AX) emerged as the optimal balance between the number of variables and the predictive performance. This finding underscores that flour composition variables alone are sufficient to predict WA, without the need for additional technological variables.

Moreover, our study highlights the significant role of minor components, such as AX, in influencing water absorption. The contribution of AX stems particularly from the water-extractable fraction (WE-AX) and polymer size, influenced by both genotype and environmental factors captured in our sample set.

Additionally, soluble starch (SS) has been identified as a novel criterion for evaluating damaged starch content, showing promise in predicting water absorption and complementing traditional assessment of damaged starch (SD).

Further work is in progress to deepen the understanding of the impact of minor components variability, particularly arabinoxylans (AX) on the rheological properties of wheat flour dough during bread-making. Investigating their natural variability represents a relevant step in this prospect.

Funding sources

This work is part of the EVAGRAIN project funded by the French national research agency with the reference ANR-20-CE21-0008.

Author contributions

Laura Rezette: Investigation, Data analysis, Performed the experiments on minor components, Visualization, Writing – original draft. Kamal Kansou: Methodology, Data analysis, Supervision, Writing – review & editing, Guy Della Valle: Supervision, Writing – review & editing Sophie Le Gall: Methodology, Supervision, Writing – review & editing, Luc Saulnier: Project administration, Methodology, Supervision, Funding acquisition, Writing – review & editing.

All authors read and approved the final manuscript.

Acknowledgements

The authors express their gratitude to Pascal Millart and Bérengère Marais from INRAE, UR 1268, Biopolymers, Interactions & Assemblies, F-44316 Nantes, France, for their assistance with biochemical analyses. Special thanks are extended to Marion Didier, Michelle Viau, and Alice Kermarrec for their invaluable help in devising the lipid analysis protocol and for their enriching contributions to advancing knowledge in this field. This work is part of the Evagrain project funded by the French National Research Agency and coordinated by the INRAE research unit BIA. The authors also thank everyone from the 10 academic and private partners who contributed to collecting the dataset on the 150 wheat samples, with particular appreciation to Marie-Hélène Morel (UMR 1208 IATE, Univ Montpellier, INRAE, L'Institut-Agro Montpellier, F-34060 Montpellier, France) for producing and sharing the protein data, Benoît Méléard (Arvalis, Institut du végétal, F-91720 Boigneville, France) for the alveograph and hardness measurements, Laurent Linossier (Limagrain Céréales Ingrédients F-63200 RIOM, France) for water absorption, and Baptiste Chambrey (Axiane Meunerie, F-35330 Val d'Anast, France) for the damaged starch and gluten measurements.

Part of the work was realized on the INRAE-BIA BIBS facility (http://www.bibs.inrae.fr/, UR1268 BIA, IBiSA, Biogenouest, Phenome-Emphasis-FR ANR-11-INBS0012, PROBE and CALIS French Research infrastructures).

Data availability

The data used in this study are available at the following address: https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=b2bada99-2f49-4e24-a0a4-c90bfa657361

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4 in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Andersson, R., Hämäläinen, M., & Åman, P. (1994). Predictive Modelling of the Bread-making Performance and Dough Properties of Wheat. *Journal of Cereal Science*, *20*(2), 129–138. https://doi.org/https://doi.org/10.1006/jcrs.1994.1053

Barak, S., Mudgil, D., & Khatkar, B. S. (2013). Relationship of gliadin and glutenin proteins with dough rheology, flour pasting and bread making performance of wheat varieties. *LWT - Food Science and Technology*, *51*(1), 211–217. https://doi.org/10.1016/J.LWT.2012.09.011

Biliaderis, C. G., Izydorczyk, M. S., & Rattan, O. (1995). Effect of arabinoxylans on bread-making quality of wheat flours. *Food Chemistry*, *53*(2), 165–171. https://doi.org/https://doi.org/10.1016/0308-8146(95)90783-4

Courtin, C. M., & Delcour, J. A. (2002). Arabinoxylans and Endoxylanases in Wheat Flour Bread-making. *Journal of Cereal Science*, *35*(3), 225–243. https://doi.org/10.1006/jcrs.2001.0433

De La Roche, I. A., Pomeroy, M. K., & Andrews, C. J. (1975). Changes in Fatty Acid Composition in Wheat Cultivars of Contrasting Hardiness'. *Cryobiology*, *12*, 5–12.

Dexter, J. E., Preston, K. R., Martin, D. G., & Gander, E. J. (1994). The effects of protein content and starch damage on the physical dough properties and bread-making quality of Canadian durum wheat. In *Journal of Cereal Science* (Vol. 20, Issue 2, pp. 139–151).

https://doi.org/10.1006/jcrs.1994.1054

Dhaka, V., & Khatkar, B. S. (2015). Effects of Gliadin/Glutenin and HMW-GS / LMW-GS Ratio on Dough Rheological Properties and Bread-Making Potential of Wheat Varieties. *Journal of Food Quality*, *38*(2), 71–82. https://doi.org/10.1111/jfq.12122

Dodds, N. J. H. (1971). Damaged Starch Determination in Wheat Flours in Relation to Dough Water Absorption. *Starch - Stärke*, *23*(1), 23–27. https://doi.org/10.1002/star.19710230108

Dowell, F. E., Maghirang, E. B., Pierce, R. O., Lookhart, G. L., Bean, S. R., Xie, F., Caley, M. S., Wilson, J. D., Seabourn, B. W., Ram, M. S., Park, S. H., & Chung, O. K. (2008). Relationship of bread quality to kernel, flour, and dough properties. *Cereal Chemistry*, *85*(1), 82–91. https://doi.org/10.1094/CCHEM-85-1-0082

Dragan, Z., Aleksandra, T., Jasna, M., Desimir, K., & Nevena, D. (2012). Relation among different parameters of damaged starch content, falling number and mechanical damage level. *Ratarstvo i Povrtarstvo*, *49*(3), 282–287. https://doi.org/10.5937/ratpov49-2345

Englyst, H. N., & Cummings, J. H. (1988). Improved Method for Measurement of Dietary Fiber as Non-Starch Polysaccharides in Plant Foods. *Journal of AOAC INTERNATIONAL*, *71*(4), 808–814. https://doi.org/10.1093/jaoac/71.4.808

Fincher, G. B., & Stone, B. A. (1974). A water-soluble arabinogalactan-peptide from wheat endosperm. *Australian Journal of Biological Science*, *27*, 117–132.

Gebruers, K., Dornez, E., Boros, D., Fraś, A., Dynkowska, W., Bedo, Z., Rakszegi, M., Delcour, J. A., Courtin, C. M., Bedő, Z., Rakszegi, M., Delcour, J. A., & Courtin, C. M. (2008). Variation in the Content of Dietary Fiber and Components Thereof in Wheats in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry*, *56*(21), 9740–9749. https://doi.org/10.1021/jf800975w

Golea, C. M., Codină, G. G., & Oroian, M. (2023). Prediction of wheat flours composition using fourier transform infrared spectrometry (FT-IR). *Food Control*, *143*(July 2022). https://doi.org/10.1016/j.foodcont.2022.109318

Greer, E. N., & Stewart, B. A. (1959). The water absorption of wheat flour: Relative effects of protein and starch. *Journal of the Science of Food and Agriculture*, *10*(4), 248–252. https://doi.org/10.1002/JSFA.2740100409

Haney, M. A. (1985). The differential viscometer. II. On-line viscosity detector for size-exclusion chromatography. *Journal of Applied Polymer Science*, *30*(7), 3037–3049. https://doi.org/10.1002/app.1985.070300725

Hoebler, C., Barry, J. L., David, A., & Delort-laval, J. (1989). Rapid Acid Hydrolysis of Plant Cell Wall Polysaccharides and Simplified Quantitative Determination of Their Neutral Monosaccharides by Gas-Liquid Chromatography. *Journal of Agricultural and Food Chemistry*, *37*(2), 360–367. https://doi.org/10.1021/jf00086a020

Jelaca, S. L., & Hlynka, I. (1971). Water-binding capacity of wheat flour crude pentosans and their relation to mixing characteristics of dough. *Cereal Chemistry*, *48*, 211–222.

Loosveld, A., Maes, C., van Casteren, W. H. M., Schols, H. A., Grobet, P. J., & Delcour, J. A. (1998). Structural variation and levels of water-extractable arabinogalactan-peptide in European wheat flours. *Cereal Chemistry*, *75*(6), 815–819.

Mangan, D., Szafranska, A., McKie, V., & McCleary, B. V. (2016). Investigation into the use of the amylase SD assay of milled wheat extracts as a predictor of baked bread quality. *Journal of Cereal Science*, *70*, 240–246. https://doi.org/10.1016/j.jcs.2016.06.015

Marion, D., Dubreil, L., & Douliez, J.-P. (2003). Functionality of lipids and lipid-protein interactions in cereal-derived food products. *Oléagineux, Corps Gras, Lipides*, *10*(1), 47–56. https://doi.org/10.1051/ocl.2003.0047

Marion, D., & Saulnier, L. (2020). Minor components and wheat quality: Perspectives on climate changes. *Journal of Cereal Science*, *94*(May), 103001. https://doi.org/10.1016/j.jcs.2020.103001

Meuser, F., & Suckow, P. (1986). Non-starch polysaccharides. *Chemistry and Physics of Baking*, 42–61.

Morel, M. H., Dehlon, P., Autran, J. C., Leygue, J. P., & Bar-L'Helgouac'H, C. (2000). Effects of temperature, sonication time, and power settings on size distribution and extractability of total wheat flour proteins as determined by size-exclusion high-performance liquid chromatography. In *Cereal Chemistry* (Vol. 77, Issue 5, pp. 685–691). https://doi.org/10.1094/CCHEM.2000.77.5.685

Morrison, W. R. (1988). Lipids in cereal starches: A review. *Journal of Cereal Science*, *8*(1), 1–15. https://doi.org/10.1016/S0733-5210(88)80044-4

Nishiyama, J., & Kuninori, T. (1987). Reactivity of sulfbydryls in reduced gluten with lipid hydroperoxides. *Agricultural and Biological Chemistry*, *51*(5), 1291–1297. https://doi.org/10.1080/00021369.1987.10868211

Pareyt, B., Finnie, S. M., Putseys, J. A., & Delcour, J. A. (2011). Lipids in bread making: Sources, interactions, and impact on bread quality. *Journal of Cereal Science*, *54*(3), 266–279. https://doi.org/10.1016/j.jcs.2011.08.011

Park, S. H., Bean, S. R., Chung, O. K., & Seib, P. A. (2006). Levels of protein and protein composition in hard winter wheat flours and the relationship to breadmaking. *Cereal Chemistry*, *83*(4), 418–423. https://doi.org/10.1094/CC-83-0418

Prabhasankar, P., & Haridas Rao, P. (1999). Lipids in Wheat Flour Streams. *Journal of Cereal Science*, *30*(3), 315–322. https://doi.org/10.1006/jcrs.1999.0289

Rakszegi, M., Bognár, Z., Li, Z., Békés, F., Láng, L., & Bedo, Z. (2010). Effect of milling on the starch properties of winter wheat genotypes. *Starch/Staerke*, *62*(2), 115–122. https://doi.org/10.1002/star.200900166

Rouau, X., El-Hayek, M. L., & Moreau, D. (1994). Effect of an Enzyme Preparation Containing Pentosanases on the Bread-making Quality of Flours in Relation to Changes in Pentosan Properties. *Journal of Cereal Science*, *19*(3), 259–272. https://doi.org/10.1006/JCRS.1994.1033

Sapirstein, H., Wu, Y., Koksel, F., & Graf, R. (2018). A study of factors influencing the water absorption capacity of Canadian hard red winter wheat flour. *Journal of Cereal Science*, *81*, 52–59. https://doi.org/10.1016/j.jcs.2018.01.012

Saulnier, L. (2019). CHAPTER 4. Types and Functionality of Polysaccharides in Cereal Grains. In T. Beta & M. E. Camire (Eds.), *Food Chemistry, Function and Analysis No. 6 Cereal Grain-based Functional Foods: Carbohydrate and Phytochemical Components* (pp. 54–84). The Royal Society of Chemistry. https://doi.org/10.1039/9781788012799-00054

Schou Jødal, A.-S., & Larsen, K. L. (2021). Investigation of the relationships between the alveograph parameters. *Scientific Reports |*, *11*(5349). https://doi.org/10.1038/s41598-021-84959-3

Selga, L., Andersson, A. A. M., Moldin, A., & Andersson, R. (2023). Determining levels of water-extractable and water-unextractable arabinoxylan in commercial Swedish wheat flours by a high-throughput method. *Journal of Cereal Science*, *109*(November 2022), 103608. https://doi.org/10.1016/j.jcs.2022.103608

Tara, K. A., Bains, G. S., & Finney, P. L. (1972). Damaged Starch and Protein Contents in Relation to Water Absorption of Flours of Indian Wheats. *Starch - Stärke*, *24*(10), 342–345. https://doi.org/10.1002/STAR.19720241005

Tipples, K. H., Meredith, J. O., & Holas, J. (1978). Factors affecting farinograph and baking absorption. II. Relative influence of flour components. *Cereal Chemistry*, *55*, 655–660.

Wang, Q., Li, L., & Zheng, X. (2020). A review of milling damaged starch: Generation, measurement, functionality and its effect on starch-based food systems. *Food Chemistry*, *315*, 126267. https://doi.org/10.1016/j.foodchem.2020.126267

Welch, R. W. (1977). A micro-method for the estimation of oil content and composition in seed crops. *Journal of the Science of Food and Agriculture*, *28*(7), 635–638. https://doi.org/10.1002/jsfa.2740280710

Zhang, L., van Boven, A., Mulder, J., Grandia, J., Chen, X. D., Boom, R. M., & Schutyser, M. A. I. (2019). Arabinoxylans-enriched fractions: From dry fractionation of wheat bran to the investigation on bread baking performance. *Journal of Cereal Science*, *87*(February), 1–8. https://doi.org/10.1016/j.jcs.2019.02.005

Figure captions

Fig. 1. *Simple linear regression between: a) Water Absorption (WA) and Protein content (Prot); b) WA and damaged starch measured by iodine absorption (SD); c) WA and Soluble Starch (SS); d) SD and SS; e) SS and Hardness; f) SD and Hardness. Dots are coloured transparently so that any overlapping ones can be discerned.*

Fig. 2. *Results of the Bayesian Information Criterium (BIC) test with composition variables. Each row represents a model, each column represents a variable. The models studied, i.e without multi-collinearity are highlighted in blue and identified (M_1, M_2 …) on the right.*

Figure caption: variable included in the model; variable included in the model but collinear with another variable from the model (VIF score >5)

WEAX: Water-Extractable Arabinoxylan content; AXTOT: Total Arabinoxylan content; A.X.WE: Arabinose to Xylose ratio from WEAX; A.X.TOT: Arabinose to Xylose ratio from AXTOT; IV.AX: Intrinsic Viscosity of WEAX; SV.AX: Specific Viscosity of WEAX; Prot: Protein content; UPP: Unextractable Polymeric protein; GluT: Total Glutenin content; GluS: Soluble Glutenin content; Gli.GluT: Gliadin to GluT ratio; Gli.GluS:

Gliadin to GluS ratio; SS: Soluble Starch; SD: Damaged starch measured by iodine absorption; C16.TOT : Total palmitic acid C16 content; C18.TOT: Total stearic acid C18 content; C181n7.TOT: Total vaccenic acid C18:1n-7 content; C181n9.TOT: Total oleic acid C18:1n-9 content; C182n6.TOT: Total linoleic acid C18:2n-6 content; C183n3.TOT: Total alpha-linolenic acid C18:3n-3 content; C16.NS : Non-Starch palmitic acid C16 content; C181n7.NS: Non-Starch vaccenic acid C18:1n-7 content; C182n6.NS: Non-Starch linoleic acid C18:2n-6 content; C183n3.NS: Non-Starch alpha-linolenic acid C18:3n-3 content.

Fig. 3. *Comparison of the models studied for predicting WA according to their BIC values and number of variables.*

Fig. 4. *Comparison of the models M_0, M_2, M_7 and M_{H3} for their predicted Water Absorption (WA) by K-fold cross validation versus the measured WA for a single repeat of the K-fold cross validation. Dots are coloured transparently so that any overlapping ones can be discerned.*

$M_0 = \beta_0+\beta_1$**Prot**$+\beta_2$**SD**;

$M_2 = \beta_0+\beta_1$**Prot**$+\beta_2$**SS**$+\beta_3$**SV.AX**$+\beta_4$**SD**;

$M_7 = \beta_0+\beta_1$**Prot**$+\beta_2$**SD**$+\beta_3$**SS**$+\beta_4$**SV.AX**$+\beta_5$**A.X.WE**$+\beta_6$**C18.TOT**$+\beta_7$**C182n6.NS**$+\beta_8$**IV.AX**$+\beta_9$**UPP**;

$M_{H3} = \beta_0+\beta_1$**Prot**$+\beta_2$**SD**$+\beta_3$**Hardness**$+\beta_4$**SS**$+\beta_5$**SV.AX**$+\beta_6$**A.X. WE**$+\beta_7$**C16.NS**$+\beta_8$**C18.TOT**$+\beta_9$**UPP**

A.X.WE: Arabinose to Xylose ratio from Water-Extractable Arabinoxylans; IV.AX: Intrinsic Viscosity of Water-Extractable Arabinoxylans; SV.AX: Specific Viscosity of Water-Extractable Arabinoxylans; Prot: Protein content; UPP: Unextractable Polymeric protein; SS: Soluble Starch; SD: Damaged starch measured by iodine absorption; C18.TOT: Total stearic acid C18 content; C182n6.TOT: Total linoleic acid C18:2n-6 content; C16.NS : Non-Starch palmitic acid C16 content.

Fig. 5. *results of the BIC test with composition and technological quality variables. Each row represents a model, each column represents a variable. A model is made up of the shaded variables. The models without multi-collinearity are highlighted and their correspondence is shown on the right.*

Figure caption: variable included in the model; variable included in the model butcollinear with another variable from the model (VIF score >5)

WEAX: Water-Extractable Arabinoxylan content; AXTOT: Total Arabinoxylan content; A.X.WE: Arabinose to Xylose ratio from WEAX; A.X.TOT: Arabinose to Xylose ratio from AXTOT; IV.AX: Intrinsic Viscosity of WEAX; SV.AX: Specific Viscosity of WEAX; Prot: Protein content; UPP: Unextractable Polymeric protein; GluT: Total Glutenin content; GluS: Soluble Glutenin content; Gli.GluT: Gliadin to GluT ratio; Gli.GluS: Gliadin to GluS ratio; SS: Soluble Starch; SD: Damaged starch measured by iodine absorption; C16.TOT : Total palmitic acid C16 content; C18.TOT: Total stearic acid C18 content; C181n7.TOT: Total vaccenic acid C18:1n-7 content; C181n9.TOT: Total oleic acid C18:1n-9 content; C182n6.TOT: Total linoleic acid C18:2n-6 content; C183n3.TOT: Total alpha-linolenic acid C18:3n-3 content; C16.NS : Non-Starch palmitic acid C16 content; C181n7.NS: Non-Starch vaccenic acid C18:1n-7 content; C182n6.NS: Non-Starch linoleic acid C18:2n-6 content; C183n3.NS: Non-Starch alpha-linolenic acid C18:3n-3 content; W.Gluten: Wet Gluten; D.Gluten: Dry Gluten; HFN: Hagberg Falling Number.

Fig. 6. *HPSEC Chromatograms of two samples 7 and 132 with low and high specific viscosity (SV.AX), respectively. The peak integration area for water-extractable arabinoxylans $WEAX_{HPSEC}$ is highlighted in yellow. Both samples have equivalent concentration of water-extractable arabinoxylans $WEAX_{HPSEC}$ as shown by RI signal with dotted lines, sample 132 have a higher intrinsic viscosity IV.AX than sample 7 resulting in higher specific viscosity as shown by DP signal in solid lines. SV.AX=IV.AX* $WEAX_{HPSEC}$.*

Table list

Table 1: Technological characteristics and flour composition obtained from the 144 sample

	Technological variables							
	WA	**Hardness**	**HFN**	**W.Gluten**	**D.Gluten**	**P/L***	**W***	**Ie***
Unit	%	/	seconds	g/100g	g/100g	/	10^{-4} J	%
Min-Max	50.4 – 63	18 – 100	190 – 462	10.5 – 40.6	3.5 – 13.4	0.13 – 2.54	68 – 546	32.8 – 73
Mean	55.9	66	348	26.4	8.4	0.83	229	54.7
CV (%)	4.6	27.9	18.7	19.2	19.4	50.6	36.6	15.7
	Flour composition variables							
	Prot	**UPP**	**GluT**	**GluS**	**Gli.GluT**	**Gli.GluS**	**SS**	**SD**
Unit	g/100 g	/	g/100 g	g/100 g	/	/	g/100 g	UCDc
Min-Max	6.1 – 14.0	0.21 – 0.60	2.48 – 5.50	1.30 - 2.77	0.86 – 1.46	1.41 – 2.71	1.11 – 6.21	8.7 – 23.9
Mean	9.4	0.46	3.87	2.07	1.08	2.01	3.22	17.5
CV (%)	16.7	14.1	18.0	14.9	10.5	13.0	31.2	14.9
	AX-TOT	**WE-AX**	**A.X.TOT**	**A.X.WE**	**IV.AX**	**SV.AX**	**C16.TOT**	**C18.TOT**
Unit	g/100 g	g/100 g	/	/	dL/g	mV.mL	g/100 g	g/100 g
Min-Max	1.27 – 2.69	0.26 – 0.86	0.56 – 0.89	0.44 – 0.71	4.14 – 8.18	201 - 834	0.24 – 0.36	0.008 – 0.027
Mean	1.83	0.50	0.72	0.60	5.86	432	0.29	0.015
CV (%)	15.5	25.7	8.5	9.2	13.4	31.3	8.75	22.0
	C181n7.TOT	**C181n9.TOT**	**C182n6.TOT**	**C183n3.TOT**	**C16.NS**	**C181n7.NS**	**C182n6.NS**	**C183n3.NS**
Unit	g/100 g	g/100 g	g/100 g	g/100 g	g/100 g	g/100 g	g/100 g	g/100 g
Min-Max	0.007 – 0.015	0.073 – 0.176	0.42 – 0.89	0.020 – 0.070	0.088 – 0.158	0.004 – 0.013	0.31 – 0.55	0.016 – 0.040
Mean	0.010	0.106	0.62	0.039	0.116	0.006	0.40	0.024
CV (%)	14.0	18.4	15.2	25.1	12.3	18.4	12.1	15.2

HFN: Hagberg Falling Number; W.Gluten: Wet Gluten; D.Gluten: Dry Gluten; P/L: Tenacity to extensibility ratio; W: Baking strength; Ie: Elasticity Index; Prot: Protein content; UPP: Unextractable Polymeric protein; GluT: Total Glutenin content; GluS: Soluble Glutenin content; Gli.GluT: Gliadin to GluT ratio; Gli.GluS: Gliadin to GluS ratio; SS: Soluble Starch; SD: Damaged starch measured by iodine absorption; AX-TOT: Total Arabinoxylan content; WE-AX: Water-Extractable Arabinoxylan content; A.X.TOT: Arabinose to Xylose ratio from AX-TOT; A.X.WE: Arabinose to Xylose ratio from WE-AX; IV.AX: Intrinsic Viscosity of WE-AX; SV.AX: Specific Viscosity of WE-AX; C16.TOT : Total C16 content; C18.TOT: Total C18 content; C181n7.TOT: Total C18:1n-7 content; C181n9.TOT: Total C18:1n-9 content; C182n6.TOT: Total C18:2n-6

content; C183n3.TOT: Total C18:3n-3 content; C16.NS : Non-Starch C16 content; C181n7.NS: Non-Starch C18:1n-7 content; C182n6.NS: Non-Starch C18:2n-6 content; C183n3.NS: Non-Starch C18:3n-3 content.

*Complementary technological variables, not used in modelling.

Table 2: Summary of models tested to predict WA with standardized variables and their corresponding coefficients and statistical metrics for prediction. All the terms of the models are statistically significant (p-value<0.05).

	Proteins variables		Starch variables		Arabinoxylans variables			Lipids variables				Statistical metric values		
	Prot	UPP	SD	SS	SV.AX	A.X.WE	IV.AX	C18.TOT	C182n6.NS	C16.NS	Hardness	n*	R^2	BIC score
Primary models														
M_{SD}			0.55									1	0.31	-43
M_{Prot}	0.63											1	0.40	-63
M_{SS}				0.72								1	0.52	-96
Usual models														
$M_{0.1}$			0.10	0.66								2	0.53	-92
$M_{0.2}$	0.44			0.58								2	0.69	-156
M_{0}	0.68		0.60									2	0.76	-192
$M_{0.3}$	0.60		0.45	0.21								3	0.78	-197
Composition models														
M_{X}	0.74		0.57		0.22							3	0.81	-218
M_{1}	0.53			0.61	0.38							3	0.83	-232
M_{2}	0.61		0.28	0.38	0.31							4	0.85	-251
M_{3}	0.64		0.28	0.41	0.25	-0.14						5	0.86	-258
M_{4}	0.61		0.28	0.40	0.24	-0.14		0.09				6	0.87	-261
M_{5}	0.63		0.34	0.35	0.22	-0.15		0.13	-0.11			7	0.88	-268
M_{6}	0.68	-0.10	0.32	0.36	0.25	-0.15		0.13	-0.12			8	0.89	-273
M_{7}	0.65	-0.09	0.35	0.35	0.29	-0.14	-0.09	0.13	-0.12			9	0.89	-273
Models with all variables														
M_{H}											0.77	1	0.59	-117

M_{H1}	0.59		0.25	0.32	0.23	-0.16			0.18	6	0.87	-265
M_{H2}	0.56		0.32	0.22	0.19	-0.18	0.12	-0.14	0.22	8	0.90	-280
M_{H3}	0.61	-0.12	0.30	0.23	0.22	-0.17	0.12	-0.15	0.24	9	0.90	-289

*number of variables included in the model. Prot: Protein content; UPP: Unextractable Polymeric protein; SS: Soluble Starch; SD: Damaged starch measured by iodine absorption; SV.AX: Specific Viscosity of Water-Extractable Arabinoxylans; A.X.WE: Arabinose to Xylose ratio from Water-Extractable Arabinoxylans; IV.AX: Intrinsic Viscosity of Water-Extractable Arabinoxylans; C18.TOT: Total C18 content; C182n6.NS: Non-Starch C18:2n-6 content; C16.NS : Non-Starch C16 content.

Additional data

S1: Selection procedure of composition variables for the study

S2: Heatmap with correlations between all variables studied to model the WA on the 144 variables

Declaration of interests

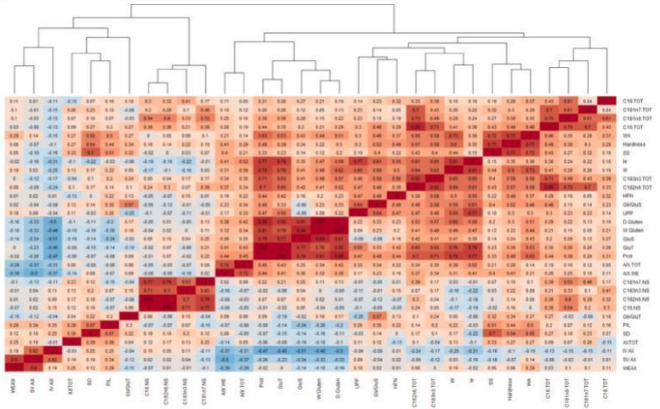
☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

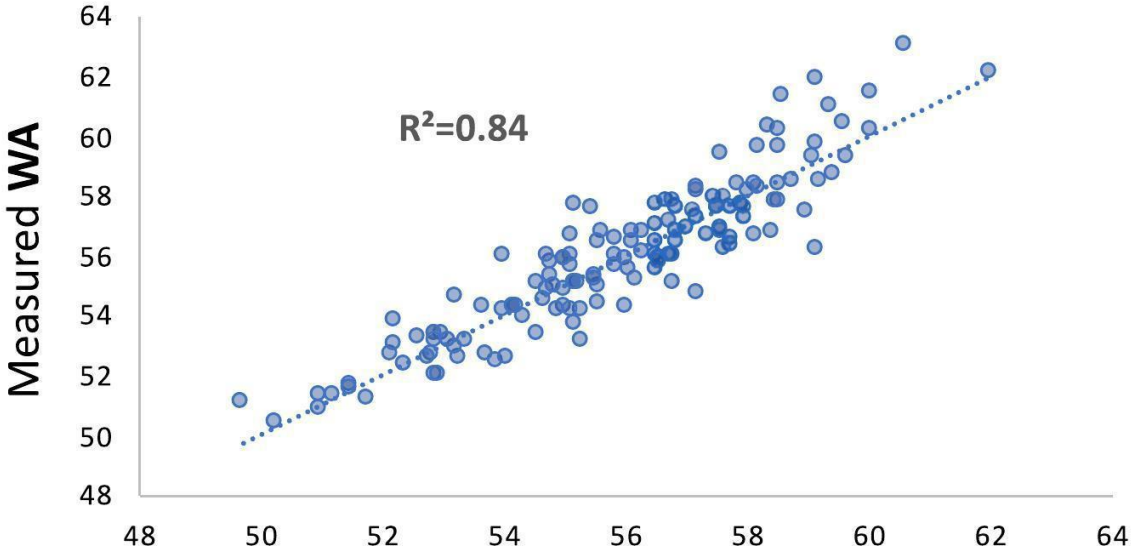
Graphical abstract

Highlights

- Protein and damaged starch are major components affecting flour water absorption
- Arabinoxylans, though minor components, significantly impact water absorption
- Water-extractable arabinoxylan induces viscosity, impacting water absorption
- Soluble starch is proposed as a marker for damaged starch



WA = water absorption
Prot= protein content
SS = soluble starch
SD = damaged starch
SV.AX = specific viscosity WE-AX



Predicted **WA** = $\beta_0+\beta_1$**Prot**$+\beta_2$**SS**$+\beta_3$**SV.AX**$+\beta_4$**SD**

Graphics Abstract

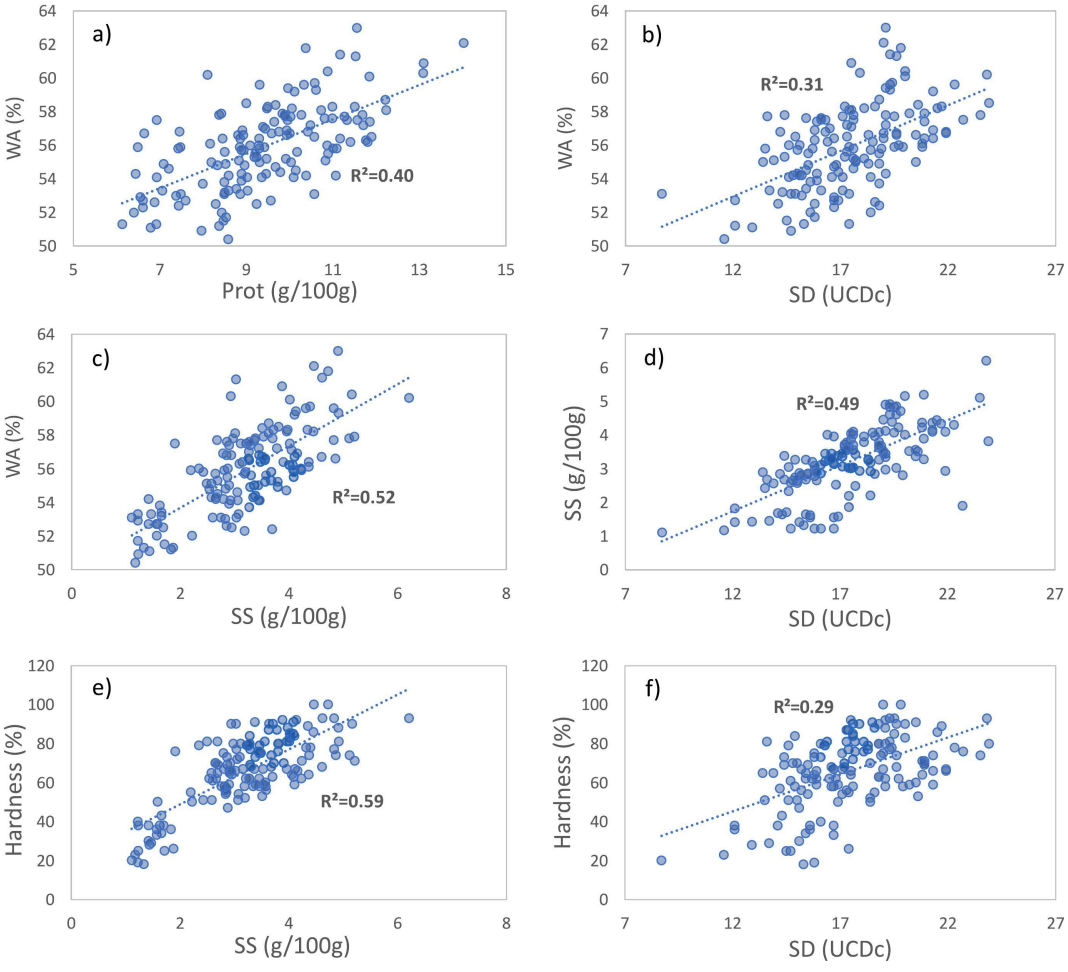

a)
WA (%)
Prot (g/100g)
R²=0.40
b)
WA (%)
SD (UCDc)
R²=0.31
c)
WA (%)
SS (g/100g)
R²=0.52
d)
SS (g/100g)
SD (UCDc)
R²=0.49
e)
Hardness (%)
SS (g/100g)
R²=0.59
f)
Hardness (%)
SD (UCDc)
R²=0.29

Figure 1

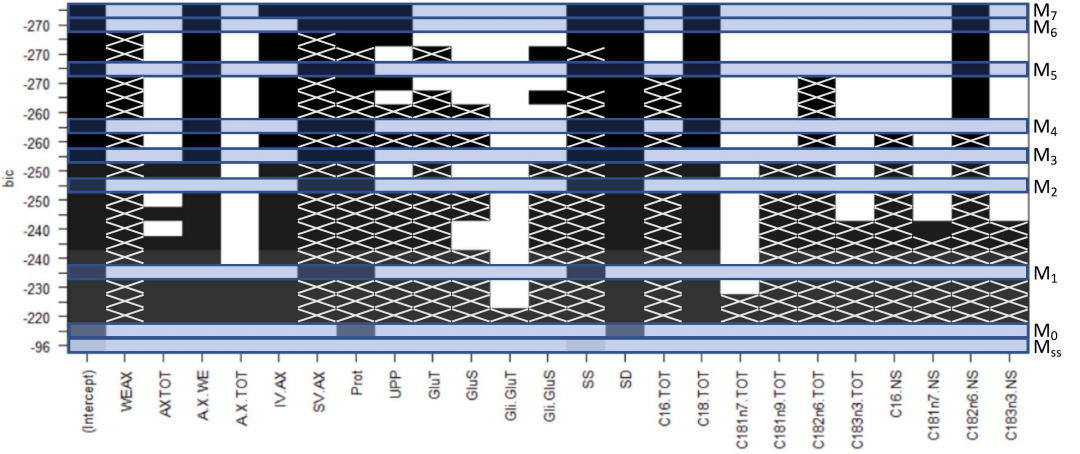

bic
-270
-270
-270
-260
-260
-250
-250
-240
-240
-230
-220
-96
M7
M6
M5
M4
M3
M2
M1
M0
Mss
(Intercept)
WEAX
AXTOT
A.X.WE
A.X.TOT
IV.AX
SV.AX
Prot
UPP
GluT
GluS
Gli.GluT
Gli.GluS
SS
SD
C16.TOT
C18.TOT
C181n7.TOT
C181n9.TOT
C182n6.TOT
C183n3.TOT
C16.NS
C181n7.NS
C182n6.NS
C183n3.NS

Figure 2

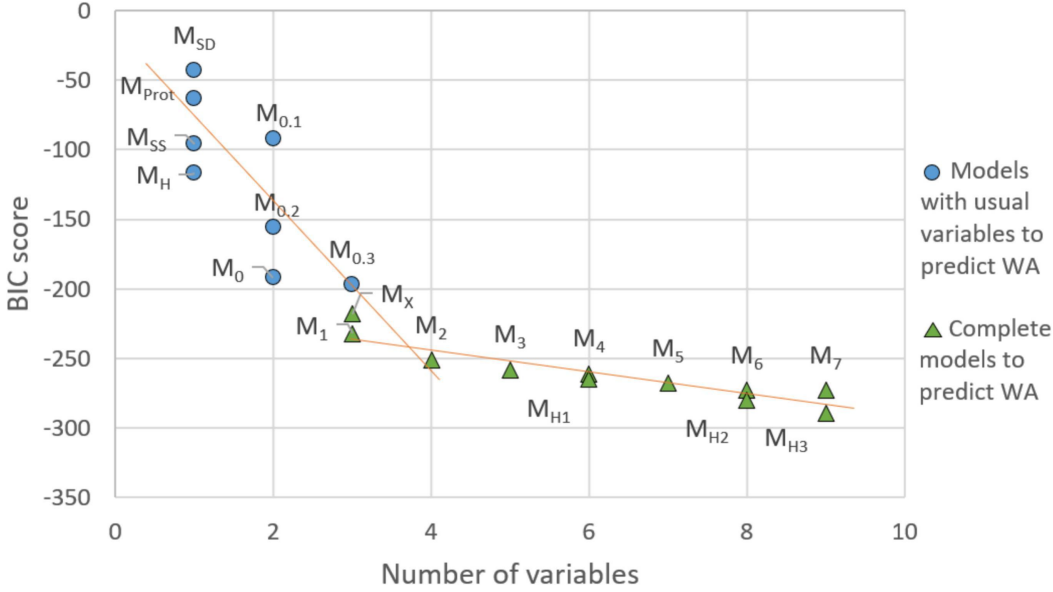

0
-50
-100
-150
-200
-250
-300
-350
BIC score
0
2
4
6
8
10
Number of variables
M_{SD}
M_{Prot}
M_{SS}
M_H
$M_{0.1}$
$M_{0.2}$
M_0
$M_{0.3}$
M_X
M_1
M_2
M_3
M_4
M_5
M_6
M_7
M_{H1}
M_{H2}
M_{H3}
Models with usual variables to predict WA
Complete models to predict WA

Figure 3

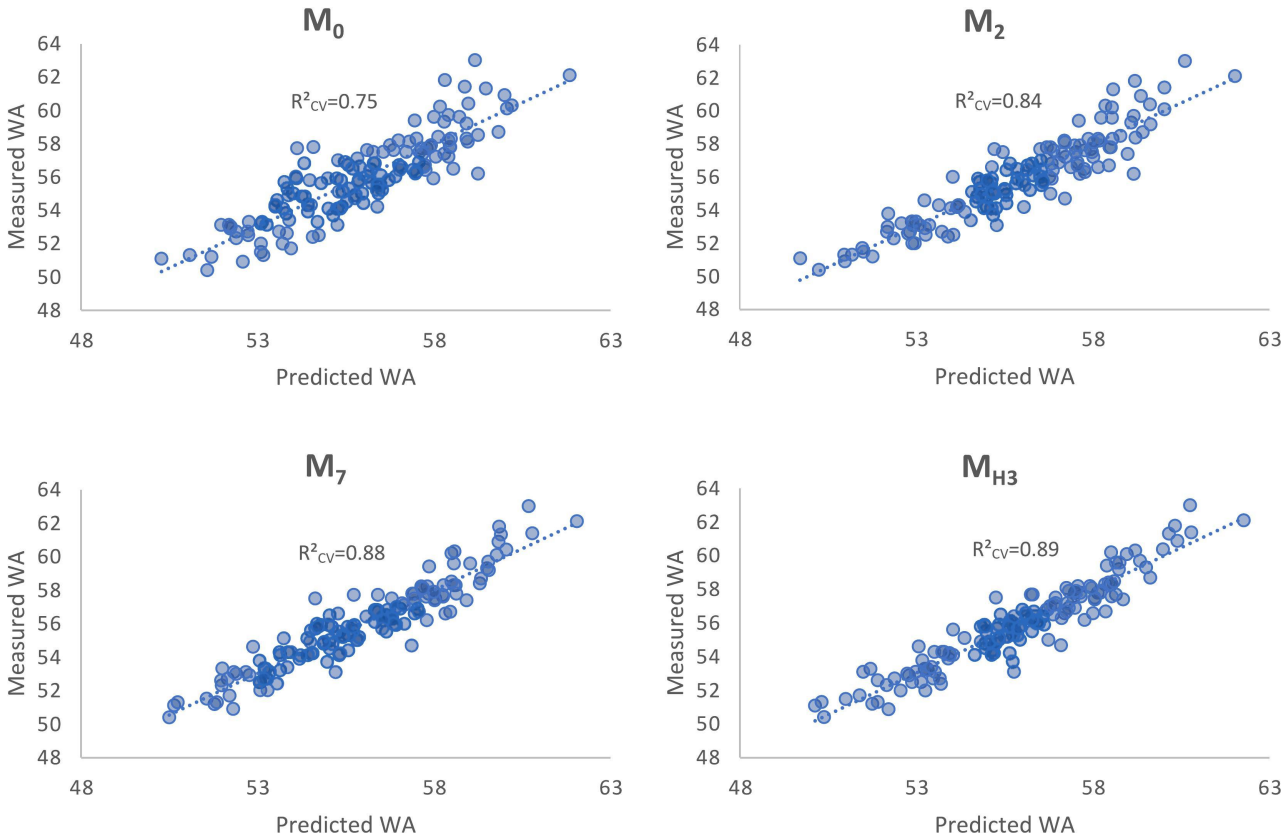

M0
R²CV=0.75
M2
R²CV=0.84
M7
R²CV=0.88
MH3
R²CV=0.89
Measured WA
Predicted WA
48
50
52
54
56
58
60
62
64
53
63

Figure 4

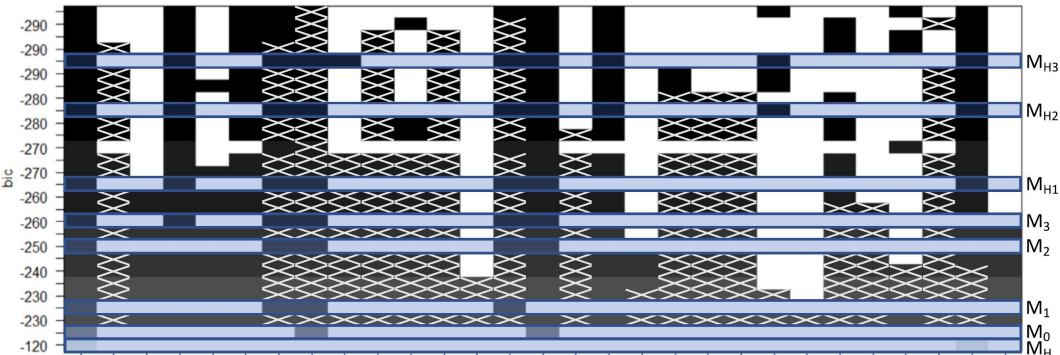

bic
-290
-290
-290
-280
-280
-270
-260
-260
-250
-240
-230
-230
-120
M_{H3}
M_{H2}
M_{H1}
M_3
M_2
M_1
M_0
M_H
(Intercept)
WEAX
AXTOT
A.X.WE
A.X.TOT
IV.AX
SV.AX
Prot
UPP
GluT
GluS
Gli.GluT
Gli.GluS
SS
SD
C16.TOT
C18.TOT
C181n7.TOT
C181n9.TOT
C182n6.TOT
C183n3.TOT
C16.NS
C181n7.NS
C182n6.NS
C183n3.NS
W.Gluten
D.Gluten
Hardness
HFN

Figure 5

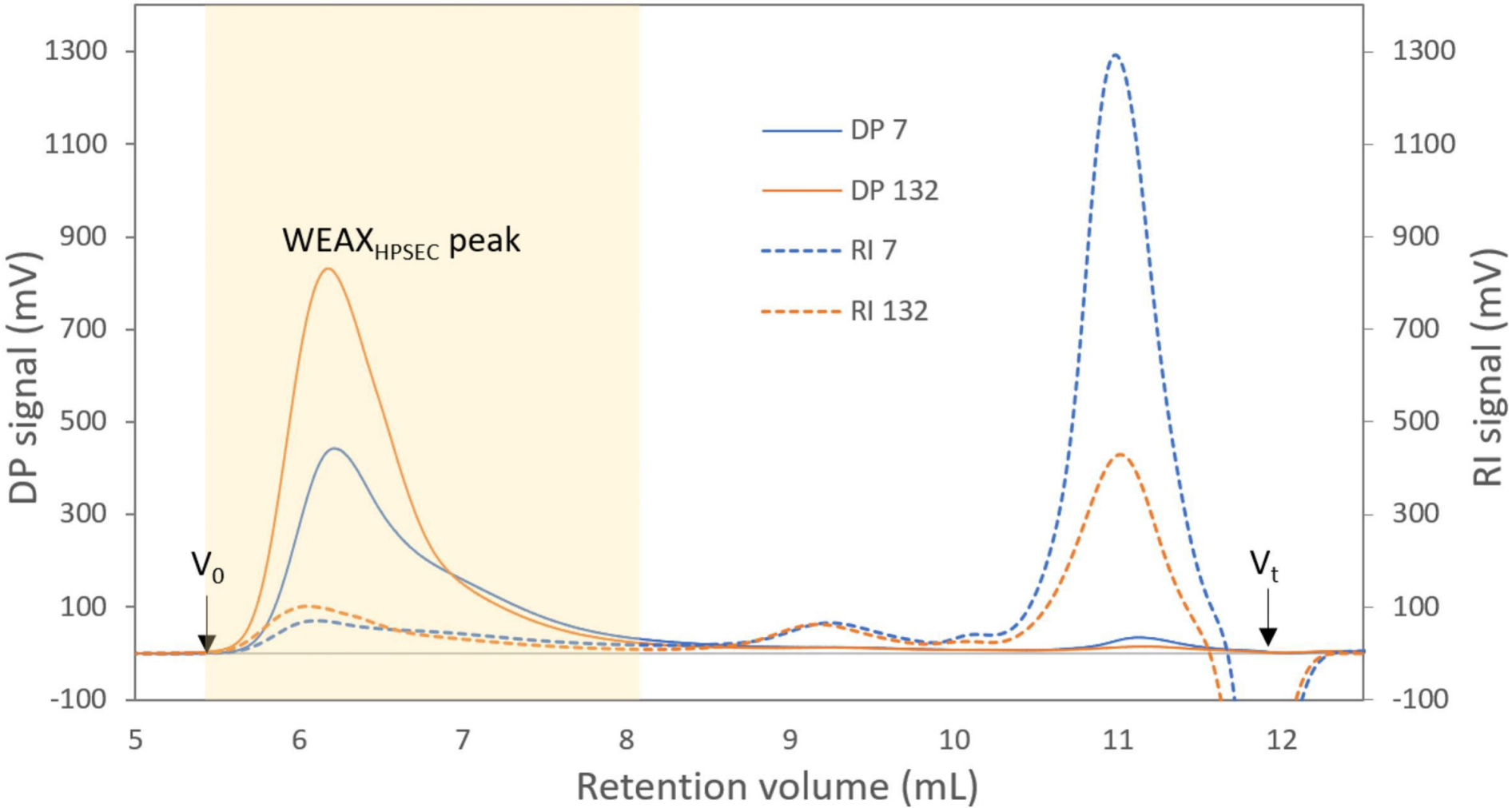

1300
1100
900
700
500
300
100
-100
DP signal (mV)
RI signal (mV)
WEAX$_{HPSEC}$ peak
DP 7
DP 132
RI 7
RI 132
V_0
V_t
5
6
7
8
9
10
11
12
Retention volume (mL)

Figure 6