Seed Science Research

cambridge.org/ssr

Machine learning links seed composition, glucosinolates and viability of oilseed rape after 31 years of long-term storage

Manuela Nagel[1], Katharina Holstein[2], Evelin Willner[1] and Andreas Börner[1]

[1]Genebank Department, Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK Gatersleben), Seeland, Germany and [2]Fraunhofer Institute for Factory Operation and Automation (IFF), Magdeburg, Germany

Research Paper

Cite this article: Nagel M, Holstein K, Willner E, Börner A (2018). Machine learning links seed composition, glucosinolates and viability of oilseed rape after 31 years of long-term storage. *Seed Science Research* **28**, 340–348. https://doi.org/10.1017/S0960258518000259

Received: 12 November 2017
Accepted: 1 June 2018
First published online: 12 July 2018

Keywords:
artificial neural networks; *Brassica napus*; fatty acids; gene bank; germination; multivariate regression; seed conservation; seed longevity; seed viability

Author for correspondence:
Manuela Nagel, Email: Nagel@ipk-gatersleben.de

Abstract

Seed longevity is influenced by many factors, a widely discussed one of which is the seed lipid content and fatty acid composition. Here, linear and non-linear regressions based on machine learning were applied to analyse germinability and seed composition of a set of 42 oilseed rape (*Brassica napus* L.) accessions grown under the same single environment and at the same time following a period of up to 31 years storage at 7°C. Mean viability was halved after 27.0 years of storage, but this figure concealed a major influence of genotype. There was also wide variation with respect to fatty acid composition, particularly with respect to oleic, α-linolenic, eicosenoic and erucic acid. Linear regression (r_L) revealed significant correlation coefficients between normal seedling appearance and the content of α-linolenic acid (+0.52) and total oil (+0.59). Multivariate regression using artificial neural networks including a radial basis function (RBF), a multilayer perceptron (MLP) and a partial least square (PLS) recognized underlying structures and revealed high significant correlation coefficients (r_M) for oil content (+0.87), eicosenoic acid (+0.75), stearic acid (+0.73) and lignoceric acid (+0.97). Oil content or a combination of oleic, α-linolenic, arachidic, eicosenoic and eicosadienoic acids and glucosinolates resulted in highest model fitting parameters R^2 of 0.90 and 0.88, respectively. In addition, the glucosinolate content, predominantly in the Brassicaceae family and ranging from 4.6 to 79.5 µM, was negatively correlated with viability ($r_L = -0.43$). Summarizing, oil content, some fatty acids and glucosinolates contribute to variations in average half-life (15.2 to 50.7 years) of oilseed rape seeds. In contrast to linear regression, multivariate regression using artificial neural networks revealed high associations for combinations of parameters including underestimated minor fatty acids such as arachidic, stearic and eicosadienoic acids. This indicates that genetic and seed composition factors contribute to seed longevity. In addition, multivariate regressions might be a successful approach to predict seed viability based on fatty acids and seed oil content.

Introduction

The concept of machine learning was introduced in the late 1960s as a means to detect patterns in data (Shalev-Shwartz and Ben-David, 2014). This field of computational statistics aims for a deeper exploration of datasets by clustering data into groups or solving a classification or (multivariate) regression tasks.

Depending on the complexity of the dataset, different approaches are used to solve regression tasks. For a simple regression analysis, a linear model might be already sufficient to describe the dataset completely and obtain valid predictions. If the dataset is more complex, e.g. various inputs that are not independent from each other, non-linear approaches have shown to be superior to describe the underlying principles between input and output data. The aim of a multivariate regression analysis is to predict the output behaviour based on the dataset correctly. For such cases, the input data are a set of numerical features, e.g. metabolite or substrate concentrations, chemical compounds or viability data (Andre, 2003; Hall, 2011; Rivas-Ubach *et al*., 2012) and is referenced as a set of feature vectors. These input vectors can be based on different units and may be continuous or discrete. Therefore, input data are normalized to cope with varying magnitudes and specific properties. The desired output vector can be a discrete quantity or continuous (Worley and Powers, 2013), e.g. a correlation value between zero and one, where larger values indicate higher correlations. Note, using only one input vector in a multivariate regression analysis still differs from a standard linear regression analysis as the correlation between input and output vector is a non-linear function. Next to the multivariate regression analysis, clustering is an unsupervised approach and the principal component (PCA) or linear discriminant analysis (LDA) are well-known examples. In PCA or LDA, data are divided into groups based on a similarity measure in the dataset (Rojas, 1996). In contrast, classification of data is usually a supervised approach, where a

CAMBRIDGE UNIVERSITY PRESS

desired output to input data is known in advance. This enables mapping of inputs to outputs to assign data of unknown classes to one of the given input classes. Thus, class membership in a classification task is a binary decision, e.g. seeds are viable or not.

Machine learning offers a variety of different algorithms to solve complex regression tasks. The simplest one would be a partial least squares (PLS) algorithm, where the sum of square residuals is minimized (Wold *et al.*, 2001). More complex models using non-linear relations are modelled on biological neurons, e.g. in multilayer perceptrons (MLP) the input data are propagated through multiple layers of activation functions (neurons) (Cybenko, 1989). Other artificial neural networks incorporate radial basis functions (RBF) and group similar feature vectors by approximation before regression is performed (Moody and Darken, 1989). To represent dependencies between input and output of a given dataset correctly, first, a mathematical model is chosen, then adapted during a training phase using an optimization algorithm, such as stochastic gradient descent or back propagation, and finally applied on validation dataset.

To assess the quality of the multivariate regression analysis the correlation coefficient r_M can be used to describe the relationship between the input value or matrix and the desired output. A further quality measure is the distance between the desired output and the predicted output, often referred to as R^2. The correlation coefficient r_M is restricted to values between −1 and 1, where values equal to −1 describe a negative correlation, values equal to 1 describe a positive correlation and values equal to 0 show no correlation. A quality measure $R^2 < 0$ shows that a horizontal line drawn in a correlation plot explains the correlation better than the results of the chosen model, $R^2 = 0$ represents an equal explanation as the line, and $R^2 > 0$ provides a better explanation than the horizontal line.

Multivariate regression analysis can produce false positive correlations. To validate the results from non-linear analysis, a cross validation scheme is employed. The available dataset is split randomly into multiple distinct subsets, where analysis is performed on each subset independently. Choosing a subset in a random way guarantees that the procedure does not produce good results by coincidence. Achieving high correlation coefficients for all cross-runs ensures that there is indeed a non-linear correlation for the overall dataset and not just the chosen random test sample. Another problem in machine learning algorithms is overfitting when models are perfectly adapted to the training data but the generalization of the obtained model is not feasible. Overfitting of artificial neural networks can be indicated by coefficients equal to 1 or highly disparate r and R^2 (Worley and Powers, 2013). To avoid overfitting effects, it is crucial to select model parameters appropriate to the structure and size of the available data.

Seed longevity, i.e. the ability of seeds to remain viable over certain storage periods, is determined by an intricate network of genetic and environmental factors. The genetic factors are associated with seed morphology and composition, whereas the environment affects by a combination of conditions prevailing during seed development, ripening, at harvest and during storage.

The major cause of the deterioration of seed quality over time is the oxidative stress, which results from a build-up of reactive oxygen species (Bailly, 2004; Kranner *et al.*, 2010; Waterworth *et al.*, 2015). Certain antioxidant compounds act to scavenge these molecules and thereby are able to protect lipids and proteins from degradation. A prominent such protectant in seeds is the fat-soluble tocopherol (Hwang *et al.*, 2014; Sattler *et al.*, 2004) which react with lipid peroxy and alkoxy radicals and so terminate the chain reaction of lipid peroxidation (Falk and Munné-Bosch, 2010). The assumption that oil-rich seeds are particularly sensitive to deterioration has been present for many years. Supporting evidence was originally based on the finding that auto-oxidation of polyunsaturated fatty acids produces free radicals, thereby compromising membrane integrity (Priestley and Leopold, 1979). The rate of oxidation is strongly dependent on oxygen concentration and temperature (Crapiste *et al.*, 1999). However, in general, the correlation between seed oil content and longevity has been described as weak (Nagel and Börner, 2010; Priestley *et al.*, 1985; Walters *et al.*, 2005).

Oilseed rape (*Brassica napus* L.) provides a substantial quantity of the world's vegetable oil production; about 44% of seed dry matter is oil (http://faostat.fao.org, 2015). The most abundant fatty acids present are linolenic (C18:3), linoleic (C18:2), oleic (C18:1) and erucic (22:1) acid. Other classes of compounds present are tocopherol (vitamin E), cellulose, phenolic acids, phytate and glucosinolates (Wittkop *et al.*, 2009). Seed longevity at cold storage (−18°C) is relatively low with a half-viability of about 25 years (Walters *et al.*, 2005).

Here, the longevity following long-term storage (7°C, 5.0% seed moisture content) of 42 oilseed rape gene bank accessions have been investigated. Accessions were grown in the same field, harvested in 1983 and stored at comparable storage conditions until 2014. The aims were: (1) to obtain an estimate of how long oilseed rape takes for seed viability to fall to 50% (P50, half-viability period) during storage at 7°C; (2) to compare key seed components and fatty acid composition of the stored material; and (3) to link seed viability with the content of key fatty acids and/or seed compounds by linear and non-linear correlation analyses using machine learning.

Materials and methods

Four replicate batches of 50 seeds of a set of 42 *B. napus* ssp. *napus* var. *napus* f. *biennis* accessions were tested for their ability to germinate in 2014. The accessions had last been multiplied together in a single field and experienced the same maternal environment in 1983. Fully mature seeds were harvested, cleaned and have been maintained at 7 ± 3°C and 5.0 ± 0.3% seed moisture content at the IPK gene bank Satellite Collection North (Malchow, Germany). The viability tests were conducted by laying seeds on moist filter paper, then keeping them under a 12 h photoperiod at 22°C. The proportion of normal seedlings (% NS) which emerged was counted, following the protocol recommended by ISTA (2014), while the overall proportion of germinated seed (%TG, total germination) was assessed after 7 days had elapsed. The %TG were compared with historic viability data collected from the same accessions in 1983, 1990, 1993 and 2009. For %NS, a comparison was only possible for historic data from 2009. Based on available %TG data, a probit analysis was conducted to estimate the half-viability periods (P50) for each and overall accessions. The probit germination percentage at storage time p_s was given by the expression $K_i - (1 \times \sigma^{-1})\ p_s$, where K_i represented the initial probit germination percentage and σ the standard deviation of the distribution of dead seeds in time (Ellis and Roberts, 1980). Percentage values of 100% response (three accessions) were corrected using 99.997% corresponding to 4.01 probit.

Seed fatty acid composition (% of total fatty acids present) was investigated in 2014, based on three replicate 200 mg samples of

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

37 out of the 42 accessions, each representing the progeny of three to five plants. Following Rücker and Röbbelen (1996), seed oil was extracted using petroleum benzine and triglycerides transmethylated by isooctane to fatty acid methyl esters, which were subjected to gas liquid chromatography (GC) using a polyethyleneglycol-2-nitroterephtalsacidester column. The concentrations of oil, protein, moisture and the content of glucosinolate in µmol g^{-1} dry weight (DW) were measured using near infrared reflectance spectroscopy (NIRS, Foss-NIRSystem 5000, Foss GmbH, Germany). NIRS calibration was adjusted by the Thuringian State Institute for Agriculture (TLL, Jena, Germany) and both NIRS and GC results are frequently evaluated by the Canadian Grain Commission.

Statistical analysis was carried out using routines implemented in GenStat software (VSN International, 2013). In particular, the data were tested for normal distribution and subjected to analysis of variance, least significant differences at $P < 0.05$ (LSD5%) and linear correlation analysis were calculated between accessions. Correlation coefficients (r_L) are only given for significant correlations at $P < 0.05$.

Non-linear statistical analysis was carried out using the neural network toolbox in MATLAB. To compute principal components, the dataset was minimized to a complete set. Here, not all data points could be provided for all accessions over a storage period of 31 years, e.g. %TG for 1990 was not available for CR 743, CR 818 and CR 822. These accessions were left out of PCA and further non-linear analyses. To determine associations between seed viability and seed composition a linear regression analysis and a multivariate regression using artificial neural networks were applied. Both correlation analyses were performed between %NS and %TG from 2014 and a single compound as input vector x. Note, in mathematical definitions lower case characters (x) are used for single values or vectors and upper case characters (X) are used for matrices. In a second step, various combinations of fatty acids and/or major components and/or historic viability results (%TG for 1983, 1990, 1993 and %NS for 2009) were used as input matrix X for a multivariate regression based on artificial neural networks (Krzanowski, 2000). Traits were chosen in a biological meaningful manner: a combination of unsaturated fatty acids, oil content and glucosinolates were expected to have highest predictability for %NS 2014 and %TG 2014. Further combinations were partly based on results of the linear regression analysis. To account for variance in the dataset due to sampling, sample size, and calibrations, each input vector x was normalized using vector L2 norm (Euclidian norm):

$$\|x\|_2 = \left(\sum_i |x_i|^2\right)^{\frac{1}{2}}$$

Furthermore, to avoid biased analysis during correlation analysis, e.g. due to different concentrations or units, all inputs were standardized (S) using z scoring as follows:

$$S = \frac{x - E[x]}{\sigma(x)},$$

with $E[x]$ the expected value and $\sigma(x)$ the standard deviation of the input vector (Worley and Powers, 2013). The significance level P was set to values ≤0.01.

An RBF using 3, 5, 7 and 10 centres, a MLP using 5 or 10 in a single hidden layer, as well as 3 and 5 neurons in the first and second layer, and a PLS using 5 components were tested for best performance. All artificial neural networks were run on the same unreduced data for reasons of comparability, where RBF models failed for some input vectors (where %TG data was missing). Each mathematical model (RBF, MLP, PLS) was run using a 10-fold cross validation. For each mathematical model and their different layouts, datasets were split into a training and generalization set (ratio 0.7). Models were trained on the training dataset. Afterwards, models were validated on the generalization dataset. Comparing results for all test runs on the training, generalization and cross-validation datasets ensures a minimization of overfitting. To evaluate performance results for the multivariate regression, the correlation coefficient r_M and the fitting parameter R^2 were chosen. Here, R^2 does not correspond to the square value of r_M but describes the correlation between predicted output and input vector, i.e. a quality measure. Still, values of r_M and R^2 should be similar. Furthermore, r_M and R^2 over all cross-validations should be comparable, with a desired R^2 between 0.4 and 1. With Y_a, the desired output vector, e.g. %NS 2014, Y_p, the predicted output vector by the neural network and mean values indicated by bars, r_M is given by:

$$r = \frac{\overline{(Y_p - \overline{Y_p})(Y_a - \overline{Y_a})}}{\sqrt{\mathrm{Var}(Y_p)\mathrm{Var}(Y_a)}},$$

and R^2 is given by:

$$R^2 = 1 - \frac{\sum_i (y_{p,i} - y_{a,i})^2}{\sum_i (y_{a,i} - \overline{Y_a})^2}.$$

Next to r_M and R^2, mean values and standard deviations of these two factors were computed and checked for overall behaviour of the non-linear correlation. In addition, all cross-runs and correlation plots for all input configurations were examined. If these were found to lie in a certain range ($r_M > 0.6$, $R^2 > 0.4$), results were considered to be mathematical meaningful (Johnson and Wichern, 2007) and overfitting could be excluded.

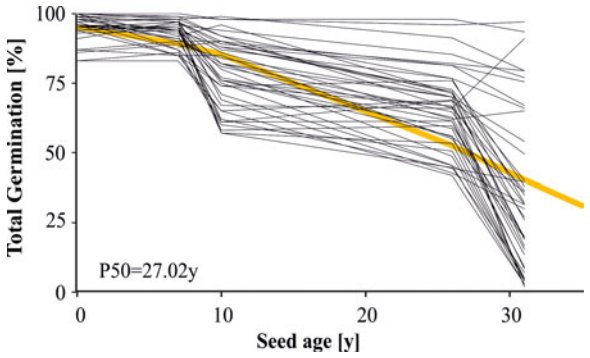

Figure 1. The decline in seed viability of oilseed rape after 31 years in long-term storage. The data are shown as the arithmetic mean derived from four replicates of 42 accessions. The curve derived from probit analysis of the data appears as a bold yellow line with the equation $y = -0.06x + 1.67$, where total germination (%) is on a probit scale. The half-viability period (P50) calculated from the probit curve is 27.02 years.

Results

The viability of freshly harvested seed in 1983 was 80–100% for the 42 accessions (mean 97.4 ± 5.4%) and between 2 and 97% by 2014 after 31 years of storage (Fig. 1, Table S1). The probit

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

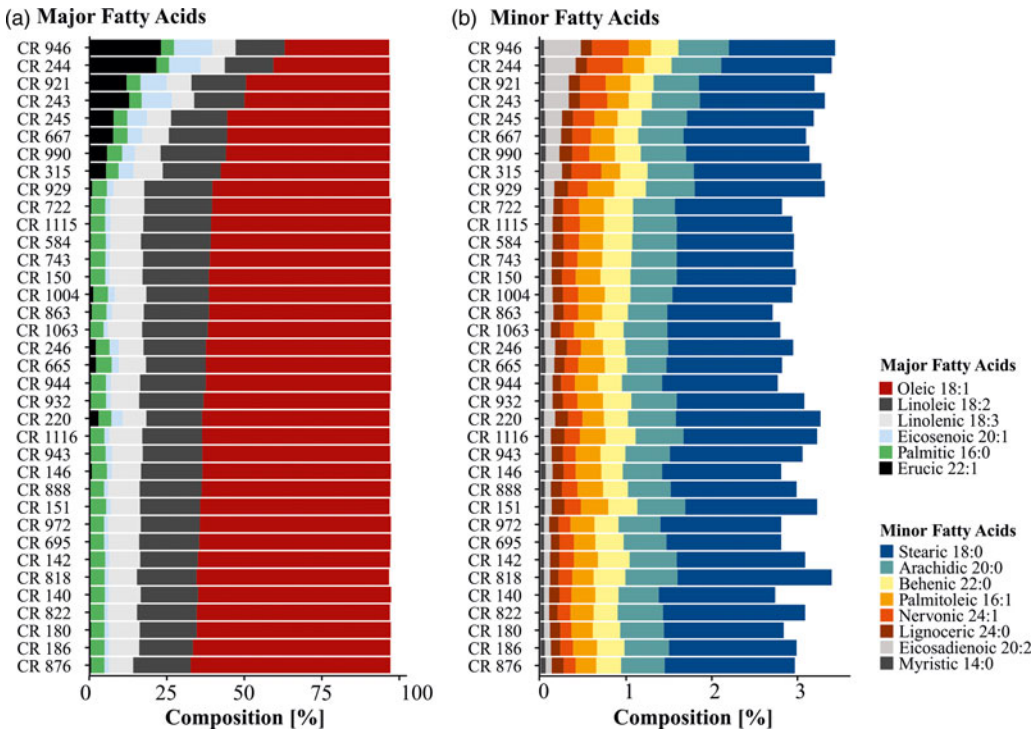

Figure 2. Fatty acid composition of 37 oilseed rape accessions measured after 31 years of storage. Fatty acid compositions are given as percentage (%) of the total fatty acid content. (a) Major fatty acids were declared when content was >4%. (b) Minor fatty acids.

analysis based on %TG of the five germination tests produced estimated half-viability periods between 15.2 and 50.7 years for an overall mean of 27.0 years. Highest correlation ($r_L = 0.71$, $P < 0.001$) was found between %TG performed in 2014 and 2009 and lowest correlation ($r_L = 0.30$, $P < 0.01$) between %TG performed in 2014 and initial germination tested in 1983 (Fig. S1). Although the accessions were submitted to the same maternal and storage environment, the low correlation coefficient between the initial and final germinability implied that factors other than initial germination were also involved.

Seed composition was significantly different ($P < 0.05$) between accessions (Table S1) and is visualised in Figs 2a,b and S2. A predominant group (29 of the 37 accessions analysed) contained high amounts of palmitic (4.7 ± 0.3%), oleic (60.4 ± 1.9%), linoleic (19.9 ± 1.4%) and α-linolenic (9.7 ± 0.9%) acids. The remaining eight accessions contained less oleic (47.1 ± 7.8%), linoleic (17.8 ± 1.8%) and α-linolenic (8.1 ± 0.7%) acids, along with more eicosenoic (7.6 ± 3.0%), eicosadienoic (0.3 ± 0.1%) and erucic (12.0 ± 6.9%) acids. The content of oil (44.6 ± 1.9%), protein (24.2 ± 1.9%) and moisture (5.0 ± 0.3%) hardly varied between accessions, while the glucosinolate content ranged from 9.1 to 85.1 μmol g^{-1} DW.

Linear regression revealed moderate associations between individual compounds and seed viability after storage (Fig. 3). Highest significant correlations were found between %NS and the contents of α-linolenic acid ($r_L = +0.52$, $P < 0.01$), oil content ($r_L = +0.59$, $P < 0.01$) and glucosinolate ($r_L = -0.61$, $P < 0.001$). Correlation coefficients were slightly lower when %TG and P50 were compared. In addition, there was an unexpected significant negative correlation between %NS and thousand seed weight ($r_L = -0.49$, $P < 0.05$).

In contrast to the linear regression, multivariate regression analysis revealed higher correlations between input vectors and seed viability. Standard approaches of machine learning were used including RBF, MLP and PLS to analyse underlying non-linear functions between the input vector or matrix and desired output vector. Networks were set up for MLP with 5, 10, or 3 and 5 neurons, PLS with 5 components, RBF with 5 centres. Due to incomplete datasets for %TG 2014, no correlation could be found for RBF networks for some compounds (Table S2). In addition, a tendency to lower R^2 for larger network set-ups could be seen for MLP. Therefore, five components were chosen for all networks and values were in the desired range of $r_M > 0.6$ and $R^2 > 0.4$ after cross-validations over all datasets (Table S3). In Tables 1, S4 and S5, superscripts indicate which network set-up produced the given results. Thereby, a high correlation coefficient r_M shows a strong relation between the input vectors and %TG 2014 or %NS 2014 whereas R^2 is a fitting parameter for the model and describes the correlation between the input vector and the predicted output vector. A negative R^2 indicates a bias in fitting of the model. The highest correlations with %TG 2014 were found with the following input vectors: lignoceric acid ($r_M = 0.97$, $R^2 = -0.03$, Fig. 4), nervonic acid ($r_M = 0.96$, $R^2 = 0.33$), glucosinolates ($r_M = 0.96$, $R^2 = 0.58$) and oil content ($r_M = 0.87$, $R^2 = 0.39$) (Table 1). Likewise, with %NS 2014, the highest correlations were found with oil content ($r_M = 0.99$, $R^2 = 0.90$), eicosenoic ($r_M = 0.99$, $R^2 = 0.62$) and stearic acid ($r_M = 0.97$, $R^2 = 0.42$) as single input vectors. Overall, correlation varied for %TG 2014 and %NS 2014 from $r_M = 0.24$ to 0.97 and from $r_M = 0.04$ to 0.99, respectively (Table 1). Correlation and R^2 increased strongly when %NS 2014 (Table S4) and %TG 2014 (Table S5) were combined with thousand seed weight (TSW) and historic viability data, respectively. In these cases, correlation ranged for %TG 2014 and %NS 2014 between $r_M = 0.74$ and 1.00 and between $r_M = 0.56$ and 1.00, respectively (Tables S4 and S5).

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

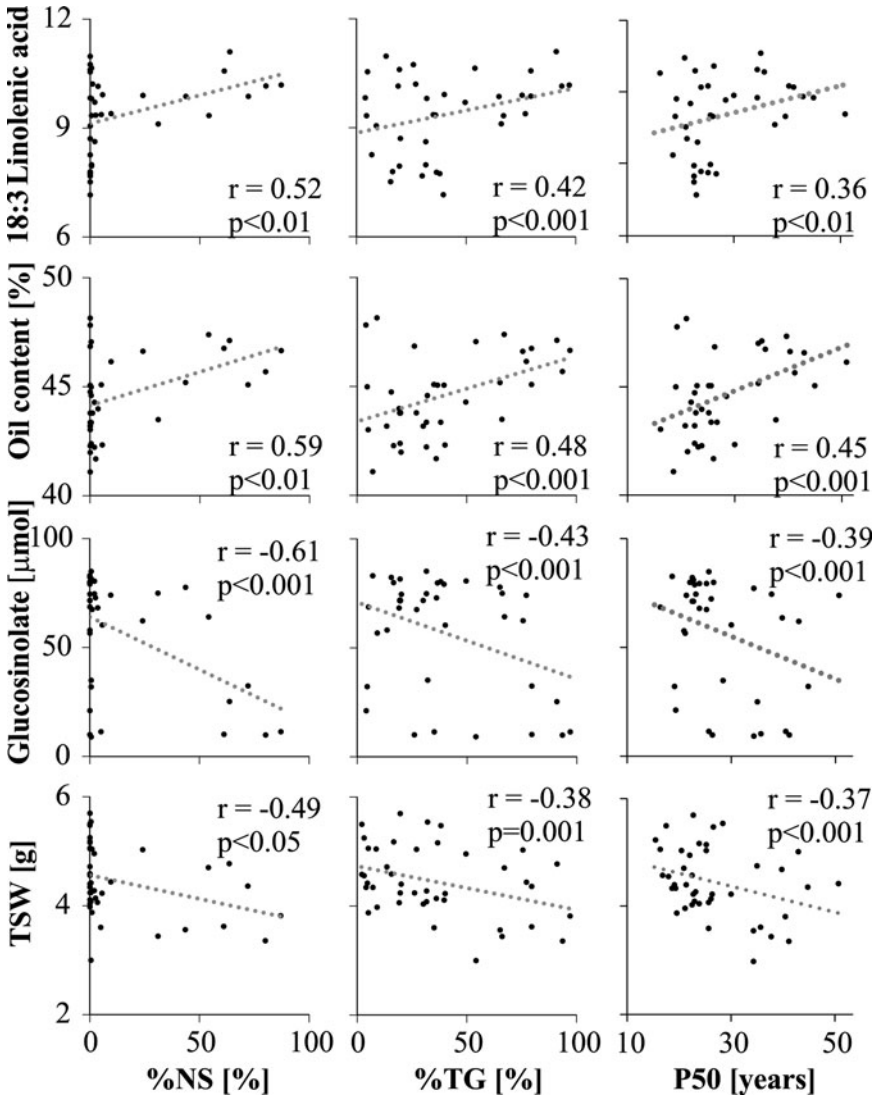

Figure 3. Significant correlations between seed viability and linolenic acid, seed components and thousand seed weight (TSW) among a set of oilseed rape accessions stored for 31 years. Seed viability is represented by normal seedlings, total germination and the half-viability period (P50).

A PCA on the complete dataset, i.e. the dataset without incomplete %TG 1990 data, provided a measure for the data structure (Fig. S3). In general, no cluster for certain accessions or the oil contents can be found. In addition to the composition performed by PCA, datasets are highlighted according to highest non-linear correlation coefficients r_M for %TG 2014 and %NS 2014. Individual P50 values, calculated for each access, were not used as an input variable because r_L between P50 and viability was not higher than the measured input parameters. Furthermore, the P50 is only an estimate of the half-viability period and standard deviation is high, and thus might lead to inaccuracies in the predictions.

Using a combination of parameters as an input matrix further improved results of the multivariate regression. The best combinations to predict seed viability after 31 years of storage (i.e % TG 2014) were: (1) myristic, stearic, oleic, α-linolenic, arachidic, eicosenoic, eicosadienoic, lignoceric and nervonic fatty acids, oil content and glucosinolates ($r_M = 0.90$, $R^2 = 0.76$); (2) α-linolenic acid, oil and glucosinolates content ($r_M = 0.84$, $R^2 = 0.56$); and (3) stearic, linoleic, arachidic, eicosadienoic, erucic fatty acids and glucosinolates ($r_M = 0.84$, $R^2 = 0.64$). The best predictive combinations for %NS 2014 were: (1) a combination of all fatty acids and all compounds ($r_M = 0.97$, $R^2 = 0.56$); (2) oleic acid, α-linolenic acid, arachidic acid, eicosenoic acid and glucosinolates ($r_M = 0.95$, $R^2 = 0.88$); and (3) myristic acid, stearic acid, oleic acid, α-linolenic acid, arachidic acid, eicosenoic acid, eicosadienoic acid, lignoceric acid and nervonic acid ($r_M = 0.96$, $R^2 = 0.85$) (Table 1). Also input matrix combinations including historic viability data (namely %TG for 1983, 1990, 1993 and %NS for 2009) and/or TSW showed a better predictability of longevity then only using the 2014 values (Fig. 4, Tables S2 and S3). By including historic viability data and TSW the fitting R^2 of the model for lignoceric acid content increased from initially −0.03 to 0.91 (Fig. 4a,b,d). In general, an input vector using only two values, e.g. TSW and myristic acid, showed high correlation values together with high R^2.

Discussion

The estimated average P50 of the current *B. napus accessions* was 27 years, which matches well with estimates obtained both from a small sample of *B. napus* (25 years, 12 accessions) and a large one of *B. oleracea* (23 years, 370 accessions) seed stored for about 40 years at −18°C (Walters *et al.*, 2005). When seeds were stored

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

Table 1. Multivariate regression reveals correlation coefficients for %NS 2014 and %TG 2014 for a combination of fatty acids and compounds

	NS% 2014		%TG 2014	
Method	r_M	R^2	r_M	R^2
Fatty acids	0.71^{x}	0.46	0.63*	0.37
Compounds	0.91^{x}	0.40	0.77^{x}	0.51
Fatty acids and compounds	0.97^{x}	0.56	0.64^{+}	−0.14
14:0 Myristic acid in %	0.57*	0.29	0.86^{+}	0.32
16:0 Palmitic acid in %	0.51^{+}	0.02	0.66^{+}	−0.04
16:1 Palmitoleic acid in %	0.04*	−0.01	0.63^{+}	0.34
18:0 Stearic acid in %	0.99^{+}	0.42	0.73^{+}	0.29
18:1 Oleic acid in %	0.88^{+}	0.25	0.60*	0.05
18:2 Linoleic acid in %	0.39^{+}	0.10	0.87^{+}	0.02
18:3 α-Linolenic acid in %	0.81*	0.32	0.76^{+}	0.16
20:0 Arachidic acid in %	0.61^{+}	0.31	0.23^{+}	0.05
20:1 Eicosenoic acid in %	0.99^{+}	0.62	0.75^{+}	0.53
20:2 Eicosadienoic acid in %	0.89^{+}	0.41	0.75^{+}	0.43
22:0 Behenic acid in %	0.22*	−0.12	0.39^{+}	0.03
22:1 Erucic acid in %	0.59^{+}	0.28	0.80^{+}	−0.02
24:0 Lignoceric acid in %	0.56^{+}	0.02	0.97*	−0.03
24:1 Nervonic acid in %	0.60*	0.02	0.91*	0.33
Oil in %	0.99^{+}	0.90	0.87*	0.39
Glucosinolates in μmol g^{-1} DW	0.73*	0.31	0.96^{+}	0.58
Protein in %	0.84*	0.24	0.80*	0.42
H_2O in %	0.63*	0.20	0.40^{+}	0.10
Stearic, linoleic, arachidic, eicosadienoic and erucic acid in %, glucosinolates in μmol	0.92^{x}	0.52	0.84^{+}	0.64
Oleic, α-linolenic, arachidic, eicosenoic and eicosadienoic acid in %, glucosinolates in μmol	0.95^{+}	0.88	0.77^{+}	0.31
α-Linolenic acid in %, oil in % and glucosinolates in μmol	0.75^{x}	0.49	0.84*	0.56
Myristic, stearic, oleic, α-linolenic, arachidic, eicosenoic, eicosadienoic, erucic, lignoceric and nervonic acid in %	0.96^{+}	0.85	0.67^{x}	0.32
Myristic, stearic, oleic, α-linolenic, arachidic, eicosenoic, eicosadienoic, erucic, lignoceric and nervonic acid in %, oil in %, glucosinolates in μmol	0.83^{x}	0.67	0.90^{x}	0.76
Myristic, stearic, oleic, α-linolenic, arachidic, eicosenoic, eicosadienoic, erucic, lignoceric and nervonic acid in %, oil, proteins and H_2O in %, glucosinolates in μmol	0.78^{+}	0.44	0.81*	0.64

Values for r_M represent best fit for all networks at $P < 0.01$ using either $^{+}$MLP with five neurons in a single layer, ^{x}RBF with five centres or *PLS with five components. Further input matrix combinations, i.e. historic viability data and thousand seed weight (TSW), can be found in Tables S2 and S3. Compounds include oil, glucosinolates, protein and H_2O content. Fatty acids include all individual fatty acids, compare Table S1.

under non-controlled ambient conditions (~20°C and 50% relative humidity), P50 of *B. napus* was only 13.9 (Priestley *et al.*, 1985) and that of *B. oleracea* only 7.3 years (Nagel and Börner, 2010). Considering that seed maternal and storage environment were comparable, a range in P50 values between 15.2 and 50.7 years suggests that genetic factors may contribute to variation in seed longevity.

Lipid composition determines oil quality and membrane structure and has profound effects on seed viability in the dry state (Hoekstra, 2005). Here, seed components varied between accessions. Contents in α-linolenic fatty acid and total oil and glucosinolates in particular were shown to correlate with seed viability after 31 years of dry storage. However, as coefficients of correlation of the linear regression were only between ±0.49 and ± 0.61, a multivariate regression using several input vectors was applied to convey complex relations between seed composition and viability. In doing so, saturated myristic, stearic, arachidic, lignoceric and unsaturated eicosenoic, eicosadienoic, oleic, α-linolenic and nervonic fatty acid were identified as predominant fatty acids influencing seed viability. Until now, very long fatty acids, e.g. lignoceric and nervonic acid, have never been discussed in relation to seed viability except in the context of human diseases and biomarkers (Lemaitre *et al.*, 2015; Clark *et al.*, 2016). Hence, multivariate regression detects relationships in complex molecular networks and enables the study of novel complex pathways.

Lipids are heterogeneously distributed across oilseed rape seed, which may facilitate different degradation processes during

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

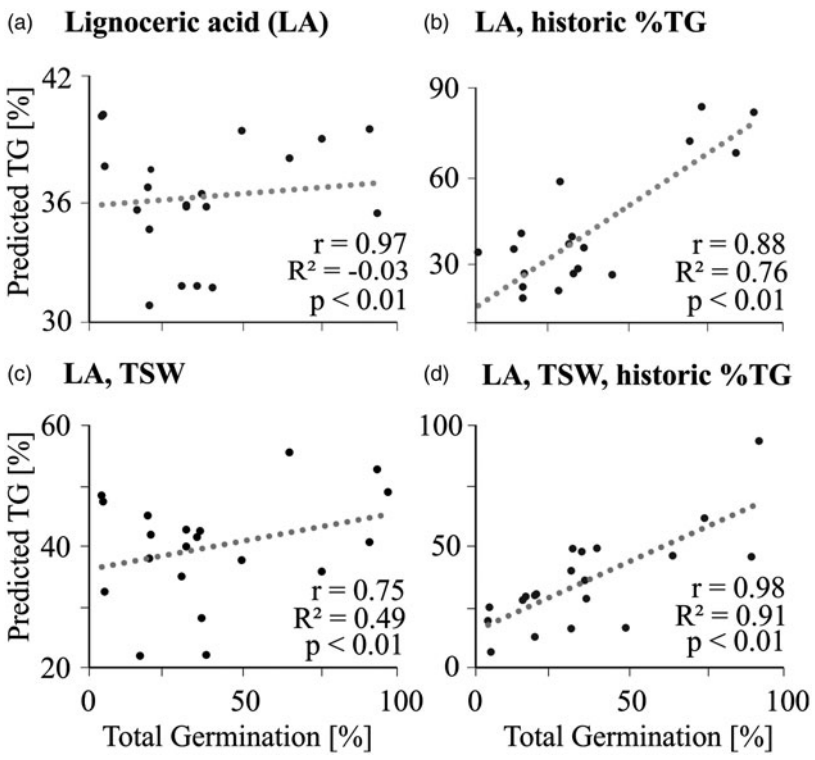

Figure 4. Prediction curve to estimate total germination (%TG) in 2014. Total germination and lignoceric acid were evaluated for the best fitted model based on multivariate regression using a partial least square (PLS) with five components. Non-linear regression was applied to (a) lignoceric acid (LA) content (%); (b) to the input matrix of lignoceric acid content (%) and historic viabilities (%TG); (c) of lignoceric acid content (%) and thousand seed weight (TSW, g); and (d) of lignoceric acid content (%), thousand seed weight (g) and historic viabilities (%TG). The non-linear regression based on artificial neural networks shows a high correlation between predicted total germination and measured total germination. Here, the importance of not only a high r but also R^2 can be seen. A higher fitting parameter ensures a better correlation between target values and actual total germination (compare improving correlation from a to d).

storage. Highest levels of palmitic acid (16:0) were found in the embryonic axis, while high levels of linoleic (18:2) and α-linolenic (18:3) acid were shown (Woodfield *et al.*, 2017) in the seed coat/aleurone layer as well as at the outer cotyledon. It may well be that the outer layers are first targets of non-enzymatic oxidation for α-linolenic acid. Indeed, there is a correlation between longevity and the average number of double bonds per polar lipid molecule (Hoekstra, 2005) which is supported by Ponquett *et al.* (1992). Although the conclusions are biased by the number of non-oily seed species, Ponquett *et al.* (1992) indicated that α-linolenic fatty acid per unit tocopherol determines the rate of oxidation. Tocopherol is a lipophilic antioxidant that is particularly abundant in oily seeds (Sattler *et al.*, 2004). Therefore, the localization as well as concentration of certain fatty acids in conjunction with antioxidants may be important in determining the rate of oxidative processes during storage.

In contrast, polyunsaturated fatty acids, such as linoleic and α-linolenic acid, are found extensively in all membranes (Harwood, 1997). Although observations were not possible here over time, Riewe *et al.* (2017) and Oenel *et al.* (2017) demonstrated in wheat and Arabidopsis that hydrolytic processes occurred during long-term storage of dry seeds. In rice, the down-regulation of different lipoxygenases reduced the production of malondialdehyde and lipid peroxides and contributed to increased tolerance against accelerated ageing (Ma *et al.*, 2015). In the present study, stearic, oleic, α-linolenic, eicosenoic, eicosadienoic, lignoceric and nervonic fatty acids, of which most are unsaturated, showed positive correlations with both %TG and %NS. It is assumed that all seed compartments are affected by oxidative processes and higher amounts of unsaturated fatty acids in particular may contribute to higher viability.

The glucosinolates found in many Brassicaceae are thought to function as a defence against herbivores. However, correlations between glucosinolate content and seed survival in the soil have never been reported (de Jong *et al.*, 2013). Linear and multivariate regressions showed that accessions accumulating a high level of glucosinolate tended to perform less well with respect to %NS. The speculation is that since glucosinolate hydrolysis products compromise bacterial membrane integrity (Borges *et al.*, 2015) it may equally damage plant cell membranes. However, antimicrobial activity facilitating glucosinolate hydrolysis is unlikely at water activity below 0.75 (Bewley *et al.*, 2013) and other mechanisms might be responsible.

Using multivariate regressions with a single input vector uncovered new significant correlations between seed composition and storability. In contrast to single input vectors, a small number of input vectors, e.g. combining one fatty acid with historic viability data or TSW (compare Fig. 4) improved predictability enormously. It seems logical to expect that historic viability data would improve r_M. However, linear regression between seed components and P50, based on historic viability data, did not increase correlation coefficients. In the seed industry, it is common to test the initial seed germination. These data may be used to estimate storability. In addition, whether TSW has an effect on seed vigour is not clear (Nagel *et al.*, 2013). In kale higher vigour was related to either bigger or smaller seeds depending on the seed lot properties (Komba *et al.*, 2007). However, oilseed rape seedlings from large seeds tend to be more vigorous and tolerant to insect damage due to a higher initial shoot biomass and higher growth rate (Bettey *et al.*, 2000; Elliott *et al.*, 2008). Therefore, there tends to be valuable information on historic viability and TSW results that supports final predictability.

High correlations for %TG 2014 and %NS 2014 can be established for quite a few network configurations. Parameters larger than five or various layers of neurons resulted in overfitting and forced the neural network to represent approximately three to four data points per centre (Johnson and Wichern, 2007). Therefore, a small network size seemed to be sufficient to convey

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press

correlations between input values and viability. However, using all available data did not always yield the best results (compare Tables S2 and S3). This would only be the case if all compounds and fatty acids were significant for the correlation. Thus, focus should be laid on set-ups where, next to high correlation r_M, high fitting parameters R^2 are also achieved. However, further studies would be interesting to follow compositional changes during storage and to apply multivariate regression for seed longevity predictions by including a higher number of accessions, thus providing a larger dataset after splitting and for cross-validation.

In conclusion, the complexity of seed longevity was shown with a unique viability dataset acquired during long-term storage for 31 years. In addition to environmental factors, seed composition was assumed to affect seed viability after long-term storage. Linear regression was shown to support this assumption but coefficients were too low to make confident interpretations. In contrast, the multivariate approaches based on machine learning were able to simultaneously analyse the impact of several parameters to reveal some important seed components, predominantly myristic, stearic, oleic, α-linolenic, arachidic, eicosenoic, eicosadienoic, lignoceric and nervonic fatty acids, oil content and glucosinolates that influence seed viability. These may be considered good candidates for developing viability prediction tools.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0960258518000259

Acknowledgements. The authors thank Veronika Miehe, Helga Schmalfeldt and Norddeutsche Pflanzenzucht Hans-Georg Lembke KG for conducting germination tests and seed composition analysis. Andrea Matros, Friedrich Melchert, Udo Seiffert and the anonymous reviewers are gratefully acknowledged for their support and helpful and critical comments on the manuscript.

Financial support. Financial support was provided by the EU [FP7 grant (311840) EcoSeed].

References

Andre M (2003) Multivariate analysis and classification of the chemical quality of 7-aminocephalosporanic acid using near-infrared reflectance spectroscopy. *Analytical Chemistry* **75**, 3460–3467.

Bailly C (2004) Active oxygen species and antioxidants in seed biology. *Seed Science Research* **14**, 93–107.

Bettey M, Finch-Savage WE, King GJ and Lynn JR (2000) Quantitative genetic analysis of seed vigour and pre-emergence seedling growth traits in *Brassica oleracea*. *New Phytologist* **148**, 277–286.

Bewley JD, Bradford KJ, Hilhorst HWM and Nonogaki H (2013) *Seeds: Physiology of Development, Germination and Dormancy*, 3rd edition. New York: Springer.

Borges A, Abreu AC, Ferreira C, Saavedra MJ, Simoes LC and Simoes M (2015) Antibacterial activity and mode of action of selected glucosinolate hydrolysis products against bacterial pathogens. *Journal of Food Science and Technology-Mysore* **52**, 4737–4748.

Clark SR, Baune BT, Schubert KO, Lavoie S, Smesny S, Rice SM, Schäfer MR, Benninger F, Feucht M, Klier CM, McGorry PD and Amminger GP (2016) Prediction of transition from ultra-high risk to first-episode psychosis using a probabilistic model combining history, clinical assessment and fatty-acid biomarkers. *Translational Psychiatry* **6**, e897.

Crapiste GH, Brevedan MIV and Carelli AA (1999) Oxidation of sunflower oil during storage. *Journal of the American Oil Chemists Society* **76**, 1437–1443.

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* (*MCSS*) **2**, 303–314.

de Jong TJ, Isanta MT and Hesse E (2013) Comparison of the crop species *Brassica napus* and wild *B. rapa*: characteristics relevant for building up a persistent seed bank in the soil. *Seed Science Research* **23**, 169–179.

Elliott RH, Franke C and Rakow GFW (2008) Effects of seed size and seed weight on seedling establishment, vigour and tolerance of Argentine canola (*Brassica napus*) to flea beetles, *Phyllotreta* spp. *Canadian Journal of Plant Science* **88**, 207–217.

Ellis RH and Roberts EH (1980) Improved equations for the prediction of seed longevity. *Annals of Botany* **45**, 13–30.

Falk J and Munné-Bosch S (2010) Tocochromanol functions in plants: antioxidation and beyond. *Journal of Experimental Botany* **61**, 1549–1566.

Hall RD (2011) Biology of plant metabolomics. *Annual Plant Reviews* **43**, 407–420.

Harwood JL (1997) Plant lipid metabolism, in Dey PM and Harborne JB (eds), *Plant Biochemistry*. San Diego, CA: Academic Press.

Hoekstra FA (2005) Differential longevities in desiccated anhydrobiotic plant systems. *Integrative and Comparative Biology* **45**, 725–733.

Hwang JE, Ahn JW, Kwon SJ, Kim JB, Kim SH, Kang SY and Kim DS (2014) Selection and molecular characterization of a high tocopherol accumulation rice mutant line induced by gamma irradiation. *Molecular Biology Reports* **41**, 7671–7681.

ISTA (2014) International Rules for Seed Testing. Bassersdorf, Switzerland: International Seed Testing Association.

Johnson RA and Wichern DW (2007) *Applied Multivariate Statistical Analysis* (6th edition). New York, Pearson Book.

Komba CG, Brunton BJ and Hampton JG (2007) Effect of seed size within seed lots on seed quality in kale. *Seed Science and Technology* **35**, 244–248.

Kranner I, Minibayeva FV, Beckett RP and Seal CE (2010) What is stress? Concepts, definitions and applications in seed science. *New Phytologist* **188**, 655–673.

Krzanowski WJ (2000) *Principles of Multivariate Analysis: A User's Perspective*. New York, Oxford University Press.

Lemaitre RN, Fretts AM, Sitlani CM, Biggs ML, Mukamal K, King IB, Song X, Djoussé L, Siscovick DS, McKnight B, Sotoodehnia N, Kizer JR and Mozaffarian D (2015) Plasma phospholipid very-long-chain saturated fatty acids and incident diabetes in older adults: the Cardiovascular Health Study. *The American Journal of Clinical Nutrition* **101**, 1047–1054.

Ma L, Zhu FG, Li ZW, Zhang JF, Li X, Dong JL and Wang T (2015) TALEN-based mutagenesis of lipoxygenase LOX3 enhances the storage tolerance of rice (*Oryza sativa*) seeds. *PLoS One* **10**), e0143877. https://doi.org/10.1371/journal.pone.0143877

Moody J and Darken CJ (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation* **1**, 281–294.

Nagel M, Behrens A-K and Börner A (2013) Effects of Rht dwarfing alleles on wheat seed vigour after controlled deterioration. *Crop and Pasture Science* **64**, 857–864.

Nagel M and Börner A (2010) The longevity of crop seeds stored under ambient conditions. *Seed Science Research* **20**, 1–12.

Oenel A, Fekete A, Krischke M, Faul SC, Gresser G, Havaux M, Mueller MJ and Berger S (2017) Enzymatic and non-enzymatic mechanisms contribute to lipid oxidation during seed aging. *Plant and Cell Physiology* **58**, 925–933.

Ponquett RT, Smith MT and Ross G (1992) Lipid autoxidation and seed ageing: putative relationships between seed longevity and lipid stability. *Seed Science Research* **2**, 51–54.

Priestley DA, Cullinan VI and Wolfe J (1985) Differences in seed longevity at the species level. *Plant Cell and Environment* **8**, 557–562.

Priestley DA and Leopold AC (1979) Absence of lipid oxidation during accelerated aging of soybean seeds. *Plant Physiology* **63**, 726–729.

Riewe D, Wiebach J and Altmann T (2017) Structure annotation and quantification of wheat seed oxidized lipids by high resolution LC-MS/MS. *Plant Physiology* **175**, 600–618.

Rivas-Ubach A, Sardans J, Pérez-Trujillo M, Estiarte M and Peñuelas J (2012) Strong relationship between elemental stoichiometry and metabolome in plants. *Proceedings of the National Academy of Sciences of the USA* **109**, 4181–4186.

Rojas R (1996) *Neural Networks: A Systematic Introduction*. Berlin: Springer-Verlag.

Rücker B and Röbbelen G (1996) Impact of low linolenic acid content on seed yield of winter oilseed rape (*Brassica napus* L.). *Plant Breeding* **115**, 226–230.

Sattler SE, Gilliland LU, Magallanes-Lundback M, Pollard M and DellaPenna D (2004) Vitamin E is essential for seed longevity, and for preventing lipid peroxidation during germination. *Plant Cell* **16**, 1419–1432.

Shalev-Shwartz S and Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press.

VSN International (2013) *GenStat for Windows* (17th edition). Hemel Hempstead, UK.

Walters C, Wheeler LM and Grotenhuis JM (2005) Longevity of seeds stored in a genebank: species characteristics. *Seed Science Research* **15**, 1–20.

Waterworth WM, Bray CM and West CE (2015) The importance of safeguarding genome integrity in germination and seed longevity. *Journal of Experimental Botany* **66**, 3549–3558.

Wittkop B, Snowdon RJ and Friedt W (2009) Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe. *Euphytica* **170**, 131–140.

Wold S, Sjöström M and Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130.

Woodfield HK, Sturtevant D, Borisjuk L, Munz E, Guschina IA, Chapman K and Harwood JL (2017) Spatial and temporal mapping of key lipid species in *Brassica napus* seeds. *Plant Physiology* **173**, 1998–2009.

Worley B and Powers R (2013) Multivariate analysis in metabolomics. *Current Metabolomics* **1**, 92–107.

https://doi.org/10.1017/S0960258518000259 Published online by Cambridge University Press