Genetic
Resources

ORIGINAL ARTICLE

Genetic Resources (2024), 5 (9), 45–60
DOI: 10.46265/genresj.DBNO8764
https://www.genresj.org
ISSN: 2708-3764

Identification of genetically plastic forms among Belarusian ancient flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) varieties using the Linum Insertion Sequence LIS-1

Maria Parfenchyk *, Valentina Lemesh, Elena Lagunovskaya, Valentina Sakovich, Andrei Buloichik, Elena Guzenko and Lyubov Khotyleva

Institute of Genetics and Cytology, National Academy of Sciences of Belarus, 27, Akademicheskaya Str, Minsk, 220072, Republic of Belarus

Abstract: The Linum Insertion Sequence 1 (LIS-1) occurs in the genetically plastic flax genotypes in response to the lack or excess of mineral and water nutrition, but also naturally, and can be transmitted to the progeny. We have analyzed 21 ancient Belarusian varieties of flax *Linum usitatissimum* convar. *elongatum* Vav. et Ell. The LIS-1 presence or absence was checked for individual plants in at minimum two generations with primer-specific polymerase chain reaction (PCR) and agarose gel electrophoresis. The studied flax varieties formed four groups: non-responsive varieties (LIS-1 was not found, group NR); responsive, which formed and completely lost the insertion (group R0); responsive, which formed and retained LIS-1 (group R1); and responsive unstable (group R2). A statistically significant difference was found in 'plant height' ($p < 0.05$), 'technical length of the stem' ($p < 0.05$) between R0 and NR, and R2 and NR LIS-1 groups. The machine learning algorithm random forest classifier was used to predict the presence, absence or heterozygosity of LIS-1 in flax plants based on their growth and reproductive characteristics. As a result, the accuracy of the prediction was 98% on test data. In terms of sources for the selection of fibre flax varieties adaptive to environmental challenges, the most promising group consists of responsive varieties that have formed LIS-1 insertion (R0, R1 and R2 groups).

Keywords: Flax, *Linum usitatissimum* convar *elongatum*, linum insertion sequence (LIS1), local varieties, machine learning, random forest classifier

Citation: Parfenchyk, M., Lemesh, V., Lagunovskaya, E., Sakovich, V., Buloichik, A., Guzenko, E., Khotyleva, L. (2024). Identification of genetically plastic forms among Belarusian ancient flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) varieties using the Linum Insertion Sequence LIS-1. *Genetic Resources* 5 (9), 45–60. doi: 10.46265/genresj.DBNO8764.

Introduction

Flax (*Linum usitatissimum* L.) has been one of the most important industrial crops for several millennia, the fibre and oil of which are used in various industries around the world (Sa *et al*, 2021). There are four convarieties of cultivated flax: *crepitans*, *elongatum*, *mediterraneum* and *usitatissimum*. Convariety *elongatum* is characterized by a plant height exceeding 70cm and side branches occupying only the upper one-third or less of the entire stem length; if the plant height falls below 70cm, stem branches are located in the upper one-fifth of the entire stem length (Diederichsen, 2019). It is common mostly in Eastern Europe (Vavilov, 1926). In Belarus, *Linum* has been one of the major technical crops for many decades. Linseed varieties are grown for oil, flax varieties are grown for fibre, and flax varieties that tend to be intermediate between oilseed and fibre types may be used to produce high oilseed yields and good-quality fibre (Rachinskaya *et al*, 2011). Flax can grow in many environments but prefers cool weather and well-drained soils with good water-holding capacity (Ehrensing, 2008).

*Corresponding author: Maria Parfenchyk (maria.parfenchyk@gmail.com)

Received: 11.07.2023 | Accepted: 07.04.2024 | Published online: 07.05.2024

Experimental studies have shown that changes in soil nutrients and water regime during flax cultivation lead to phenotypic changes, which are accompanied by genomic rearrangements (Evans *et al*, 1966; Cullis, 1976, 1981; Goldsbrough *et al*, 1981). One of the heritable genomic changes in response to nutrient and water stress is the appearance of Linum Insertion Sequence-1 (LIS-1). A more detailed study of the occurrence of LIS-1 showed that the appearance of the insertion is limited to individuals (genotypes) that respond to growth conditions by modifying their genome (Chen *et al*, 2009). Single-copy insertion LIS-1 is assembled from short sequences scattered throughout the flax genome in a short period of time before flowering (Cullis, 1976). The LIS-1 is a 5.7kb nucleotide sequence that inserts at a specific site in the flax genome. This specific site contains two genes, inhibitor of growth-1, and kip-related cyclin-dependent kinase inhibitor-2. The target sequence is also modified when LIS-1 is inserted. A total of 129 single nucleotide polymorphisms and InDels were found in the target sequence when comparing lines with and without LIS-1 (Bickel *et al*, 2012). Contrasting conditions of mineral nutrition of seedlings caused the appearance on the plastic (or responsive) genotypes (line Pl) of two types of stable genotrophs: L-genotrophs and S-genotrophs (Durrant and Nicholas, 1970; Durrant and Jones, 1971). Treatment with low or imbalanced nutrients (different concentrations of nitrogen, phosphate and potassium in soil, or lack of water) gives rise to the small, or S, genotroph. Morphologically the S-genotrophs are shorter than the Pl line, have a non-branching stem, hairy capsule septa, and contain single-copy insertion LIS-1. A high nutrient and water treatment results in the large, or L-genotrophs, which are much taller than the S-genotrophs, have more branching stems than Pl or S, hairless capsule septa, and do not contain LIS-1. Both L- and S-genotrophs have been shown to be better adapted to the nutrient environment in which they were induced. In the responsive flax genotype, which did not form stable genotrophs, LIS-1 was lost in the absence of inducing conditions (Bickel *et al*, 2012; Chen *et al*, 2009, 2005). The characteristics altered in the genotrophs include height, weight, number of branches, the presence of hairs on seed capsule septa, total nuclear DNA contents, the number of genes coding for the large ribosomal RNAs and the 5S ribosomal RNAs as well as a number of other repetitive sequence families (Chen *et al*, 2005).

Chen *et al* (2005) have shown that the insertion LIS-1 could occur also naturally in many flax and linseed varieties, and concluded that the external environment can act both as an inducer of directed genetic variability and as a selective factor for beneficial mutations. Flax is a self-pollinator, and the ability to modify the genome may be an adaptive property (Cullis, 1986).

To sum up, in responsive flax genotypes many genomic rearrangements occur in response to environmental challenges, including the occurrence of LIS-1. Using LIS-1 as a marker of genome plasticity is a fast and cost-efficient tool for primary screening of genotypes as the insertion could be detected by polymerase chain reaction. Thus, the LIS-1 sequence is a promising molecular marker for identifying flax forms with genome plasticity and, accordingly, adaptive capacity.

Artificial intelligence and machine learning algorithms are used in different fields of science to solve problems of classification or regression. By learning from existing data, supervised, semi-supervised or unsupervised techniques could be applied in chemistry (Raghunathan and Priyakumar, 2022) for predicting properties and designing molecules and materials; in the pharmaceutical industry (Volkamer *et al*, 2023) for the predictions of bio-activity and physical properties; and in active learning (Bajorath, 2022) for drug discovery. It was shown that random forest provides comparable performance and easier interpretation for many applications (Volkamer *et al*, 2023) or outperforms other models (Yang *et al*, 2022), like support vector machine, decision tree, and extreme gradient boosting tree algorithms. Random forest is a supervised ensemble method that randomly builds and integrates multiple decision trees to create a forest structure. The choice of the machine learning algorithm depends on the data structure, data types and questions you want an answer to (Hu and Xing, 2021).

The aim of this study was to investigate local ancient Belarusian flax varieties with the LIS-1 insertion as a marker of genome plasticity and, based on available morphological features, to construct a classification model to predict the presence or absence of the LIS-1 insertion using a random forest classifier.

Materials and methods

Plant material

For the analysis, 21 ancient local varieties of flax *Linum usitatissimum* convar. *elongatum* Vav. et Ell. were used. Seeds were obtained in 1998 from the N.I. Vavilov Institute of Plant Genetic Resources (VIR) genebank, where they were collected during Vavilov expeditions from 1923 to 1958 in Belarus. Plant material origin is shown in Table 1. Since 1998, these varieties have been cultivated and studied at the Biological Experimental Station of the Institute of Genetics and Cytology of the National Academy of Sciences of Belarus. In 2000, the National Bank of Seeds of Plant Genetic Resources of Belarus was established for the conservation, investigation and use of plant genetic resources (Privalov *et al*, 2021), where investigated varieties are included and conserved. Since the seed material was obtained from the VIR genebank, the accession numbers are given according to VIR.

Experimental conditions

The 21 studied local ancient varieties of flax were sown and cultivated in the Biological Experimental Station of

Table 1. Flax (*Linum usitatissimum* convar. *elongatum* Vav. et Ell.) accessions included in the study. Information includes accession numbers in the VIR collection, dates of inclusion in the collection, origin and LIS-1 group. Information about the exact location of some sampling has not been preserved.

Accession number (VIR code)	Year added	Place of origin	Belarus region	LIS-1 group
624_595	1923	Homielskaja voblasć	South-East	R2
624_596	1923	Homielskaja voblasć	South-East	NR
624_776	1923	Viciebskaja voblasć	Nord	R1
624_781	1923	Minskaja voblasć, Červieński district	Center	NR
624_784	1923	Belarus	Unknown	NR
624_786	1923	Belarus	Unknown	R1
624_787	1923	Belarus	Unknown	R1
624_789	1923	Belarus	Unknown	R1
624_791	1923	Mahiloŭskaja voblasć, Čerykaŭski district	East	R0
624_1043	1924	Viciebskaja voblasć, Połacki district	Nord-East	R0
624_1044	1924	Viciebskaja voblasć, Haradocki district	Nord	R1
624_5462	1939	Minskaja voblasć, Dokšycki district	Center	R1
624_5463	1939	Viciebskaja voblasć, Pastaŭski district	Nord-West	R2
624_5464	1958	Viciebskaja voblasć, raka Dzisna	Nord	R2
624_6213	1958	Viciebskaja voblasć, Šarkaŭščynski district	Nord-West	R2
624_6214	1958	Viciebskaja voblasć, Hłybocki district	Nord-West	R1
624_6215	1958	Hrodzienskaja voblasć, Navahrudski district	West	R1
624_6216	1958	Bresckaja voblasć	South-West	R0
624_6220	1958	Bresckaja voblasć	South-West	R2
624_6219	1958	Bresckaja voblasć, Ivanaŭski district	South-West	R2
624_6222	1958	Hrodzienskaja voblasć, Karelicki district	West	R1

the Institute of Genetics and Cytology in Minsk from 2017 to 2022. The studied varieties are self-pollinating, and no crosses were carried out with them. About 50 plants per variety were grown each year, 10 plants were taken to test the phenotype, and 5–10 plants per variety were tested by polymerase chain reaction (PCR) for the presence or absence of the LIS-1 insertion. No further studies were conducted for those varieties for which LIS-1 was not found. Seeds were collected from each plant found to have LIS-1, planted again in the following year, and the presence or absence of LIS-1 was checked again. For each variety, there were at least two generations of plants with traceable LIS-1 presence or loss.

Nitrogen fertilizer was applied to the soil in accordance with the field fertilization schedule, both when preparing the soil for seeding and during plant growth. Plants were watered as needed.

Weather conditions (rainfall, mm and temperature, °C) during the flax vegetation period from May to August for the study years are shown in Figure 1.

The conditions of the growing season varied depending on the year. The average temperatures in May, June, July and August were 12.83°C, 18.51°C, 18.48°C, and 18.52°C, respectively. The coldest years were 2017 and 2020 (mean growing season temperatures were 16.48°C and 16.20°C, respectively), and the warmest was 2018 (average temperature 18.70°C), with a 6-year average growing season temperature of 17.09°C. The wettest years were 2018 and 2019 (total rainfall during the growing season was 328 and 333mm), and the driest year was 2020 (total rainfall 240mm), with an average rainfall during the growing season over six years of 307.17mm.

The optimal temperature for flax development is from 15–18°C. The phenological stages of flax are: germination (May), leaf development (end of May), active growth (June), budding and flowering (June, July), development of seed capsules (July), ripening of seed (end of July, August). Flax continues growing until the end of flowering.

Phenotypic characterization

The morphological features of the flowers and plants of the varieties were evaluated according to the flax descriptors (Maggioni *et al*, 2001; Nôžková *et al*, 2016). For the flower, the following were detected: shape of the flower, shape of the corolla, size of the corolla, shape of the petals, longitudinal folding of the petals, colour of corolla (the petals colour, when fully developed), colour of the veins of the petals, anther colour, seed colour. For the plant: foliation, plant height, technical length of the stem; number of productive seed capsules per plant; and number of seeds in the capsule. For the modelling, morphological features included an extended list of characteristics, which were not investigated before: the ciliation of seed capsule septa and the presence of anthocyanin pigmentation in the hypocotyl in addition to the plant height, the technical length of

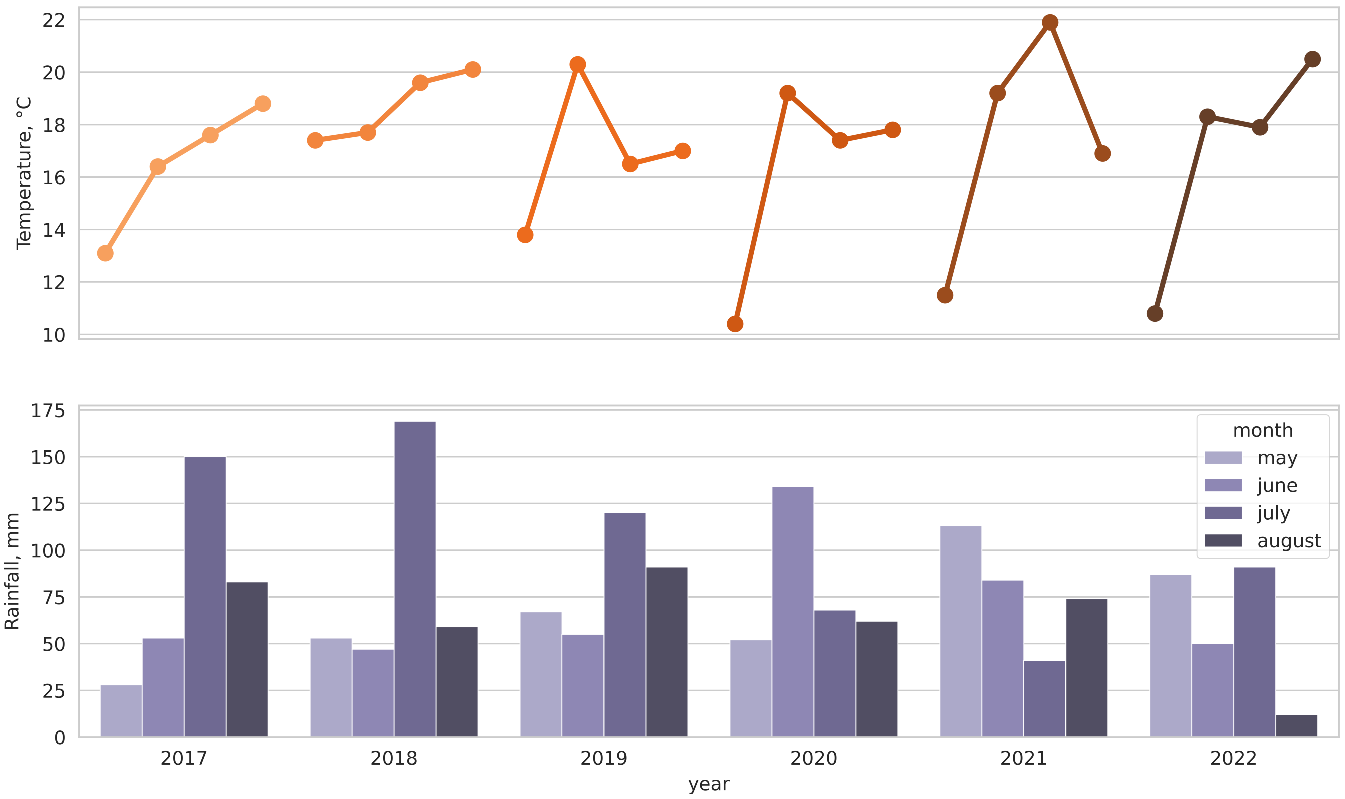

Figure 1. Rainfall (mm) and temperature (°C) in May, June, July and August for each year of the study.

the stem, the number of productive seed capsules per plant and the number of seeds in the capsule. For the modelling, seven flax varieties which were grown in 2022 were used, for which these characteristics were available: 624_6222, 624_6219, 624_1044, 624_791, 624_786, 624_6215, and 624_789. Ciliation of seed capsule septa was shown to be associated with LIS-1 presence/absence (Chen *et al*, 2005), and therefore this trait could be important for the machine learning model. We included the presence of anthocyanin pigmentation in the hypocotyl characteristic in our list of traits as one of the possible responses of flax genotypes to abiotic stress. This information is available in Supplemental Table 1.

Molecular genetic analysis

For molecular analysis, upper leaves were collected at the budding stage, since in the experiment of Cullis (1976) it was shown that the LIS-1 insertion is assembled from short fragments of DNA distributed throughout the genome in the short time preceding flowering. Leaves (50–100mg) of 5–10 individual plants per accession were taken for DNA isolation according to Sambrook and Russell (2006). Briefly, plant material was placed in 2.0ml microcentrifuge tubes, 15μl of TE buffer (10mM Tris-HCl, pH 7.5; 1mM EDTA) was added and the tissue ground using a Tissue Lyser tissue homogenizer (Qiagen). Then, 400μl of lysis solution (TrisHCl 1M, pH = 8.0, NaCl 5M, EDTA 0.5M, SDS 10%) was added to each tube and incubated for 10 minutes, with periodical shaking. Then, 600μl of phenol-chloroform mixture (1:1 v/v) was immediately added, mixed gently and centrifuged at 10,000rpm for 10 minutes. The supernatant was transferred to new tubes, and 600μl of a mixture of chloroform-isoamyl alcohol (24:1 v/v) was added, mixed and centrifuged for 2 minutes at 10,000rpm. The upper phase was transferred into clean tubes, and 800μl of ice-cold 96% ethanol was added, mixed with gentle rocking, and placed for 15 minutes at -20°C, then centrifuged for 7 minutes at 10,000rpm, and the supernatant was completely removed. The next step was to dissolve the DNA on a vortex at low speed in 100μl of a 1.2M NaCl solution. Then 300μl of ice-cold 96% ethanol was added, and the DNA was allowed to precipitate (for up to 2 hours at -20°C) and then centrifuged for 4 minutes at 12,000rpm. The precipitate was washed in 300μl of 70% cold ethanol, and the alcohol was removed with a pipette. DNA was dissolved in 100μl of double-distilled water. The DNA concentration in the resulting solution was measured using an Ultrospec 3300 pro spectrophotometer (Amersham Biosciences, USA) at a wavelength of 260nm (ultraviolet spectrum).

The LIS-1 insertion was detected by sequence-specific polymerase chain reaction (PCR) as described by Chen *et al* (2009). The primers used for amplification of LIS-1 at the insertion site were: 2 and 3' (5'-ggtttcagaactgtaacgaa-3' and 5'-gaggatggaagatgaagaagg-3'), 18 and 19' (5'-cataaattcagtcctatcgac-3' and 5' taacagctcggatctaggc 3'); the absence of the insertion was detected by amplification with primers: 2 and Pl9' 5'-ggtttcagaactgtaacgaa-3' and 5'-gcttggatttagacttggcaac-3'. The sizes of the amplified fragments were 416, 398 and 417bp, respectively (Chen *et al*, 2009). Electrophoretic separation was carried out in 1.5% agarose gel, with further detection in ultraviolet light.

Statistical analysis

Correlation analysis using Python libraries: NumPy (Harris *et al*, 2020), SciPy (Virtanen *et al*, 2020) was performed to measure the strength of the relationship between the examined features and to calculate their association.

Generalized linear models were used to identify whether there were significant contributions of interaction between factors genotype (LIS-1 groups) and weather conditions (temperature and rainfall by month, average temperature and total rainfall) to predict the dependent variables (plant height, technical length of the stem, number of seed capsules per plant, and number of seeds in the capsule). The formula with interaction was: 'dependent variable ~ LIS_group*weather_condition'. The Shapiro test was applied to check the normality of models residuals with statistical significant effect. Analysis was run in Python version 3.10.9 using the statsmodels (Seabold and Perktold, 2010).

Analysis of variances (ANOVA) was performed to identify if factor LIS-1 groups had a significant effect on dependent variables: plant height, technical length of the stem, number of seed capsules per plant, and number of seeds in the capsule. The Shapiro-Wilk test (Wickham *et al*, 2022) was used to determine the normality and Levene's test (Wickham *et al*, 2022) was used to determine homoscedastisity to check ANOVA assumptions. Shapiro test was applied to check the normality of models residuals distribution. For the 'technical length of the stem' and 'number of seed capsules per plant', non-parametric Kruskal-Wallis ANOVA test was applied as the normality assumptions failed. Post-hoc Tukey's test for parametric and Dunn's test for non-parametric distributed characteristics were used for LIS-1 groups pairwise comparisons for models with significant effect of LIS-1 group factor on morphological characteristics of plants. P-values were taken into account under Bonferroni correction. Analysis was run in R version 4.1.2 using the lme4 (v1.1-35.1; Bates *et al* (2015)), rstatix (v0.7.2; Kassambara (2023)), dplyr (v1.1.4; Wickham *et al* (2022)) packages.

Machine learning model

For the analysis, a sample of 56 plants with known morphologic characteristics were taken from seven flax varieties: 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), 624_6219 (R2 group), and 624_791 (R0 group). Before building the model, the sample of plants were tripled using the function pandas.DataFrame.sample() with replacement (pandas.pydata.org, The pandas development team (2020)). Analysis was run in Jupyter notebook (Kluyver, 2016).

The basis of the method is the use of a set of single classifiers (a decision tree) to make the final decision on the classification of objects. The selection of hyperparameters for the calculation was carried out using the GridSearchCV algorithm, which checks each combination of hyperparameters from a given range of values and selects the most successful one. The number of iterations was set to seven for better cross-validation results. The range of values was as follows: param_grid = {'bootstrap': [True, False], 'max_depth': [3, 5, 7, 10], 'max_features': ['auto', None], 'criterion': ['gini', 'entropy'], 'n_estimators': [600, 800, 1000]}. The best combination of hyperparameters was calculated as follows: {'bootstrap': True, 'criterion': 'gini', 'max_depth': 7, 'n_estimators': 1000}. Random forest classifier algorithm was run with the following options: {'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 7, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 1000, 'n_jobs': None, 'oob_score': False, 'random_state': 13, 'verbose': 0, 'warm_start': False}. The number of features used at nodes for split was defined by grid search as sqrt(n_features). The criterion for feature importance estimation is gini impurity, which shows how each feature decreases the impurity of the split.

In total, nine features were used: four numeric (plant height; technical length of the stem; number of productive seed capsules per plant; number of seeds in the capsule), and two categorical (ciliation of septa and anthocyanin pigmentation in the hypocotyl). Categorical features were encoded to numeric with the pandas library method get_dummies(), and resulted in five features (ciliation of septa as: hairless hh, heterozygous Hh, hairy HH); the presence of anthocyanin pigmentation in the hypocotyl as: week or missing). Target variable was the presence of LIS-1 insertion ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote', when three PCR products were revealed).

The dataset was divided into training and test sets, with the proportion of test size being 0.3. At first, the machine learning algorithm will use only the training part of data (70% from all dataset) including features (characteristics) and the target variable (what we want to predict). On this data, the algorithm will learn about the structure and relationships between features, and finally make a prediction about the target variable. Then the algorithm takes a new 30% of test data (only features) and makes a prediction again. When comparing the real values of the test target variable with the predicted values, we can assess the degree of success achieved. As a criterion for the success of the classification, the accuracy score was calculated (the ratio of correctly classified objects to the total number of operations performed). A confusion matrix provided a visualization of additional information about classification results: precision, recall and f1-score. Abbreviations in formulas to calculate these metrics were: TP (true positive, when both the actual and predicted values were 1), TN (true negative, when both the actual and predicted values were 0), FP (false positive, when the actual value was 0, but the predicted value was 1), FN (false negative, when the actual value was 1, but the predicted value was 0). Classification metrics were the following: precision, the proportion of positive identifications that were actually correct: TP/(TP+FP); recall, the proportion of actual positives correctly identified: TP/(TP+FN); f1-score, the weighted average of precision and recall – this score takes both false positives and false negatives into account: 2*(Recall * Precision) / (Recall + Precision).

Figures 1, 5, 6 and 7 were produced using Python with the 'matplotlib' and 'seaborn' (Hunter, 2007; Waskom, 2021) packages. Figure 3 was produced using R with the ggplot2 (v3.5.0; Wickham (2016)) package.

Results

Morphological characteristics of the studied flax varieties are shown in Table 2 and Table 3.

Varieties with the largest mean plant height were 624_6216, 624_6219, 624_6222, whereas varieties with the largest mean technical stem length were also 624_6216, 624_6219 and 624_5462. Varieties with the smallest mean plant height and technical stem length were 624_595, 624_6215, 624_781, and 624_6215, 624_786, 624_781.

The flowers differed in diameter, from small (624_782, 624_784, 624_6213, 624_6217, 624_6222) to large (624_791, 624_6216). According to the colour of the corolla and the anthers, the following combinations were noted: violet/bluish, blue/creamish, light blue/bluish, light blue/dark bluish, blue/bluish, and blue/dark bluish. Only one variety had cream-coloured anthers (624_776).

The majority of studied local ancient Belarusian varieties had a medium-sized corolla, a regular shape of the flower with circular petals, blue flower petals, anthers and veins of the petals, and brown seeds.

Presence of LIS-1

The studied flax varieties could be divided into four groups, corresponding to LIS-1 presence and preservation: R0, R1 and R2, which are responsive genotypes as they formed LIS-1 insertion, and NR group which includes non-responsive genotypes, i.e. LIS-1 insertion was not detected for them. Data is shown in Table 4. This LIS-1 group subdivision of accessions is the first attempt to generalize data.

Shared and individual morphological characteristics of the flower, seeds and stem of the studied flax varieties grouped by LIS-1 presence are shown in Figure 2. The analysis was run online using Venny (Oliveros, 2015). Based on the morphological characteristics listed in Table 3, we could not distinguish responsive and non-responsive accessions. The NR group of accessions had no distinctive individual characteristics, yet shared ten common traits with all four groups, and only one trait in common with the R1 and R2 groups (small corolla), and one with the R2 group (violet colour of petals). Groups of responsive accessions (R0, R1, R2) had three characteristics in common (blue colour of petal veins and petals, semi-star shape of flower), R0 and R1 groups had in common dark bluish colour of anthers, R0 and R2 groups were characterized by stem high foliation, and R1 and R2 groups had light brown colour of seeds coat. The R0 group had one individual characteristic (large flower); the R1 group had two individual traits (elliptical shape of petals and creamish anthers). So, we can consider that the presence or absence of LIS-1 insertion could have morphological effects, but the studied characteristics of the flower and stem are not sufficient for an exact morphological differentiation.

Average morphological characteristics of the four LIS-1 groups and two responsive groups (R0, R1, R2 are responsive, NR is non-responsive) are shown in Table 5.

Correlation analysis

We used correlation analysis to establish the relationship between quantitative characteristics, environmental conditions (rainfall, temperature) during vegetation period, and genetic group defined by LIS-1, as we want to reveal which factors impacting on morphologic characteristics depending on plant development stage and genetic group.

The LIS-1 groups exhibited significant negative correlations with both plant height and technical length of the stem (r = -0.294** and r = -0.248*, respectively, Table 6). Conversely, there was a strong positive correlation between plant height and technical length of the stem (r = 0.889***), between plant height and number of seeds per capsule (r = 0.25**), and plant height and temperature in August (r = 0.314***). Similarly, positive significant correlations were found between technical stem length and temperature in August (r = 0.427***), rainfall in June (r = 0.107*), rainfall in July (r = 0.2**), whereas negative significant correlations were observed between technical stem length and number of seed capsules per plant (r = -0.269**), temperature in June (r = -0.466***), and rainfalls in May and August (r = -0.341***, r = -0.271***). Plant height also exhibited negative significant correlations with temperatures in May and June (r = -0.08**, r = -0.264**), and rainfalls in May and August (r = -0.289***, r = -0.265***). Positive significant correlations were found between the number of seeds per capsule and the number of seeds capsules per plant (r = 0.45***), temperatures in June and the number of seeds in the capsule (r = 0.323***), rainfall in May and the number of seeds in the capsule (r = 0.368***), with additional statistically significant correlations between the number of seeds per capsule and temperatures in June (r = 0.241*) and rainfall in May (r = 0.151*). The number of seeds per

Table 2. Quantitative morphological characteristics of flax varieties. Numbers given represent the mean ± standard deviation (std) of ten plants examined.

Variety	Mean plant height ± std	Mean technical length of the stem ± std	Mean number of seed capsules per plant ± std	Mean number of seeds in the capsule ± std
624_595	53.08±1.82	40.37±1.997	7.80±1.64	7.92±1.64
624_596	47.97±1.90	36.7±2.01	7.3±1.55	7.63±1.55
624_776	53.97±2.57	42.34±2.50	6.43±1.32	7.82±1.32
624_781	48.03±2.52	36.78±2.39	7.14±1.76	7.68±1.76
624_784	55.09±2.32	43.64±2.90	7.18±1.37	7.85±1.37
624_786	53.22±2.50	39.67±2.57	8.52±2.25	8.07±2.25
624_787	55.52±2.23	42.62±2.06	6.53±1.059	8.2±1.059
624_789	56.43±2.64	45.08±2.38	7.695±1.53	8.08±1.53
624_791	55.90±3.24	43.34±3.067	7.92±2.048	8.058±2.05
624_1043	55.98±2.32	46.08±3.17	7.0017±1.45	7.985±1.45
624_1044	55.46±3.042	42.88±3.51	7.55±1.41	8.18±1.41
624_5462	57.06±4.58	48.2±4.43	6.496±1.56	7.79±1.56
624_5463	57.07±2.33	46.40±2.49	6.052±1.15	8.26±1.15
624_5464	56.42±3.15	47.22±3.08	5.64±1.107	7.705±1.11
624_6213	58.24±3.00	46.88±2.84	6.228±1.28	8.094±1.28
624_6214	53.94±1.96	43.87±2.089	6.61±1.202	7.92±1.202
624_6215	51.93±3.22	40.34±4.02	8.296±1.38	8.016±1.38
624_6216	63.86±2.99	52.76±3.22	6.53±1.27	8.32±1.27
624_6219	61.33±3.58	48.062±3.98	8.556±1.77	8.136±1.77
624_6220	54.09±3.41	45.68±2.64	6.697±1.43	7.44±1.43
624_6222	58.36±4.83	46.26±3.55	7.067±1.51	8.08±1.51

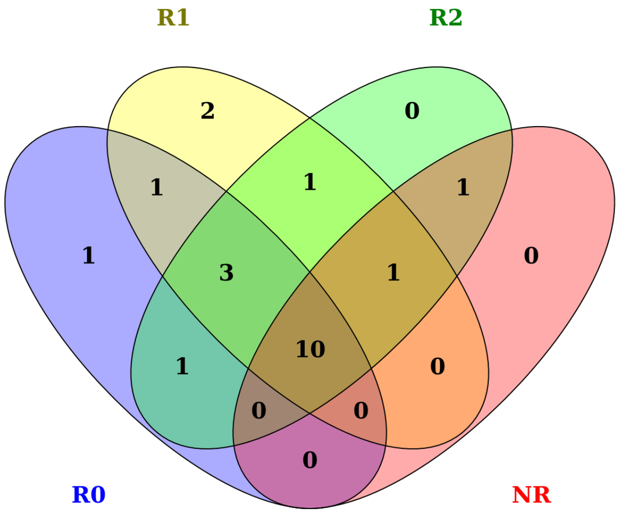

Figure 2. Venn diagram showing the number of common morphological characteristics of the flower, seeds and stem for the varieties grouped by LIS-1. R0: responsive varieties, formed LIS 1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

Table 3. Morphological characteristics of flax varieties and assignment to LIS-1 groups. The morphological characteristics of the flower, petal, foliation and seed coat colour were stable. *Anther colour in 2023, which is not included in this manuscript, for accession K-776 was not creamish. R0: responsive varieties, formed LIS-1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

Variety	Flower			Petal				Anther colour	Foliation	Seed colour	LIS-1 group
	Shape	Corolla shape	Size of corolla	Shape	Longitudinal folding	Colour of corolla	Vein colour				
624_595	Regular	Funnel	Medium	Circular	Absent	Violet	Violet	Bluish	Medium	Brown	R2
624_596	Regular	Funnel	Medium	Circular	Absent	Violet	Violet	Bluish	Medium	Brown	NR
624_776	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Creamish*	Medium	Light brown	R1
624_781	Regular	Plate like	Medium	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	NR
624_784	Regular	Funnel	Small	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	NR
624_786	Regular	Funnel	Small	Circular	Absent	Light blue	Blue	Dark bluish	Medium	Brown	R1
624_787	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_789	Semi-star	Plate like	Small	Elliptical	Absent	Blue	Violet	Bluish	Medium	Brown	R1
624_791	Regular	Plate like	Large	Circular	Absent	Light blue	Blue	Bluish	Medium	Brown	R0
624_1043	Regular	Funnel	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R0
624_1044	Regular	Plate like	Medium	Circular	Absent	Blue	Violet	Dark bluish	Medium	Brown	R1
624_5462	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_5463	Semi-star	Plate like	Medium	Circular	Absent	Light blue	Blue	Bluish	High	Brown	R2
624_5464	Semi-star	Plate like	Medium	Circular	Absent	Blue	Violet	Bluish	Medium	Light brown	R2
624_6213	Regular	Plate like	Small	Circular	Absent	Light blue	Blue	Bluish	Medium	Brown	R2
624_6214	Regular	Plate like	Medium	Circular	Absent	Light blue	Violet	Bluish	Medium	Brown	R1
624_6215	Regular	Plate like	Medium	Circular	Absent	Blue	Blue	Bluish	Medium	Brown	R1
624_6216	Semi-star	Plate like	Large	Circular	Absent	Light blue	Blue	Dark bluish	High	Brown	R0
624_6219	Regular	Plate like	Medium	Circular	Absent	Blue	Violet	Bluish	Medium	Brown	R2
624_6220	Regular	Funnel	Medium	Circular	Absent	Blue	Blue	Bluish	High	Brown	R2
624_6222	Regular	Funnel	Small	Circular	Absent	Blue	Blue	Dark bluish	Medium	Brown	R1

Table 4. Accessions and the LIS-1 groups to which they could be affiliated based on the LIS-1 insertion presence and preservation.

LIS-1 group	Definition	Accessions
R0	Responsive, formed the insertion and then, over a number of generations, completely lost it.	624_6216, 624_791, 624_1043
R1	Responsive, formed the insertion and retained it over a number of generations.	624_776, 624_5462, 624_6214, 624_6222, 624_786, 624_787, 624_789, 624_1044, 624_6215
R2	Responsive, formed the insertion and then partially lost it over a number of generations.	624_595, 624_6220, 624_5463, 624_5464, 624_6213, 624_6219
NR	Non-responsive varieties, in which the insertion was not found.	624_784, 624_596, 624_781

Table 5. Quantitative morphological characteristics of the four LIS-1 groups and two responsive groups. R0: responsive varieties, formed LIS-1 and completely lost the insertion; R1: responsive varieties that retained the insertion; R2: responsive varieties that formed the insertion and then partially lost it over a number of generations; NR: non-responsive varieties, LIS-1 insertion not detected.

LIS-1 group	Mean plant height, cm ± std	Mean technical length of stem, cm ± std	Mean no. of seed capsules per plant ± std	Mean no. of seeds in the capsule ± std
R0	59.70±2.95	48.66±3.19	7.16±1.62	8.07±1.62
R1	54.88±3.16	43.16±3.02	7.28±1.48	8.02±1.48
R2	55.63±2.90	44.72±2.82	6.82±1.40	7.88±1.40
NR	50.18±2.15	39.07±2.43	7.24±1.60	7.73±1.60
Responsive group				
Responsive (R0+R1+R2)	55.95±3.04	44.61±2.99	7.11±1.48	7.98±1.48
Non-responsive (NR)	50.18±2.15	39.07±2.43	7.24±1.60	7.73±1.60

capsule exhibited a negative significant correlation with temperature in July (r = -0.261**).

As a result, plant height and technical stem length are correlated with the LIS-1 group, requiring a thorough analysis of their relationship; correlations between environmental conditions (rainfall, temperature) and quantitative characteristics (plant height, technical length of the stem, number of seed capsules per plant, and number of seeds per capsule) align with flax preferences across various stages of plant development.

The interaction effects of temperature in June were significant for groups R0, R2 and NR, impacting plant height ($p < 0.05$), while the effect for the R1 group was not statistically significant ($p = 0.108$) when implementing the generalized linear model. Effect on the border of significance between genotype (LIS-1 groups) and rainfall in May for the reproductive trait 'number of seeds in the capsule' was observed in the NR group ($p = 0.056$).

Analysis of variances (ANOVA)

The analysis of variances (ANOVA) showed statistically significant differences among LIS-1 groups for 'plant height' (p = **0.00119**) and 'technical length of the stem' (p = **0.00581**). Results are shown in Table 7.

For traits where ANOVA showed significant results, Tukey's and Dunn's tests were implemented to reveal the statistical significance of differences among LIS-1 groups in pairwise comparisons (Table 8).

A statistically significant difference was found in the plant height between groups NR and R0 ($p < 0.01$), and NR and R2 ($p < 0.05$). Similarly, for the technical length of the stem, a statistically significant difference was observed between the same groups NR and R0, and NR and R2 ($p < 0.05$). Before the sequential Bonferroni correction, significant differences were shown for groups R1 and R0, and R1 and NR ($p < 0.05$) in the technical length of the stem.

Accessions grouped as R0 and R2 were characterized by greater mean plant height and technical length of the stem compared to accessions of the NR group. The R1 group was statistically significantly taller than the NR group ($p = 0.0365$ before the sequential Bonferroni correction) for what concerns the technical length of the stem. In contrast, for reproduction-related traits such as 'number of seed capsules per plant' and 'number of seeds in the capsule', responsive genotypes (R0, R1, R2) did not significantly differ from non-responsive genotypes (NR). Figure 3 shows comparisons of all four groups by morphological characteristics. The R0, R1 and R2 groups are generalized as 'responsive group', and NR as 'non-responsive'.

Thus, from the plant's point of view, the most important thing is to leave seeds at the end of the vegetation period. Both responsive and non-responsive genotypes, revealed by LIS-1, are successful at this. But, from the point of identifying potentially useful genotypes for fibre flax selection, LIS-1 responsive

Table 6. Correlation analysis performed between the examined features (LIS-1 genetic group, plant height, technical length of the stem, number of seed capsules per plant, number of seeds per capsule, and temperature and rainfall in May, June, July and August). *, **, ***: statistically significant at $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. Bold font indicates significant correlations between quantitative characteristics, environmental conditions, and genetic group defined by LIS-1. Correlations between environmental conditions (rainfall, temperature) were out of the scope of manuscript, so are not discussed.

		Quantitative traits			
	LIS-1 groups	**Plant height**	**Technical stem length**	**No. of seed capsules per plant**	**No. of seeds per capsule**
Plant height	**-0.294****				
Technical stem length	**-0.248***	**0.889*****			
No. of seed capsules per plant	-0.036	-0.001	**-0.269****		
No. of seeds per capsule	-0.189	**0.251****	-0.012	**0.452*****	
Temp May	-0.001	**-0.08****	0.06	-0.213	-0.21
Temp June	0.011	**-0.264****	**-0.466*****	**0.323*****	**0.241***
Temp July	0.003	-0.177	-0.008	-0.072	**-0.261****
Temp August	-0.001	**0.314*****	**0.427*****	-0.029	-0.035
Rainfall May	0.009	**-0.289*****	**-0.341*****	**0.368*****	**0.151***
Rainfall June	0.004	0.157	**0.107***	-0.132	-0.117
Rainfall July	-0.006	0.098	**0.2****	-0.224	-0.11
Rainfall August	-0.005	**-0.265*****	**-0.271*****	-0.216	-0.124

Table 7. F-values, statistics and p-values from analysis of variance (ANOVA) and non-parametric Kruskal-Wallis analysis of variance of morphological characteristics for flax accessions grouped into four LIS-1 groups. Bold font indicates a significant difference.

Morphological characteristic	**ANOVA**	**Shapiro-Wilk normality test**
Plant height	F = 5.643, p-value = **0.00119**	W = 0.98964, p-value = 0.484
Number of seeds in the capsule	F = 1.549, p-value = 0.206	W = 0.9933, p-value = 0.8254
	Kruskal-Wallis	
Technical length of the stem	statistic = 12.5, p-value = **0.00581**	-
Number of seed capsules per plant	statistic = 1.41, p-value = 0.703	-

Table 8. Tukey's and Dunn's tests results and p-values. Bold font indicates a significant difference after the sequential Bonferroni correction. Numbers in brackets indicate a significant difference before the correction.

	Morphological characteristics	
	Plant height	**Technical length of the stem**
LIS-1 groups	**Tukey's test results, p-values**	**Dunn's test, p-values**
R1_R0	0.0613309	0.223 (0.0371)
R2_R0	0.1956210	1 (0.187)
NR_R0	**0.0004668**	**0.00379** (0.000632)
R2_R1	0.9618691	1 (0.403)
NR_R1	0.0710239	0.219 (0.0365)
NR_R2	**0.0420870**	**0.0580** (0.00966)

genotypes appear to have an advantage visualized by larger plant height and stem length.

Ciliation of septa

A. Durrant and O. R. Joarder (1978) described that the differences between flax genotrophs must be due to differences in gene activity and regulation induced by the environment. In particular, the capsule quantitative character H-h (hairy-hairless septa) is conditioned by multiple genes. Also, H and h are stable in homozygotes but unstable in a heterozygous state. HH and hh, and their adjacent regulators, could be controlled by genes elsewhere in the genome. The capsule character was defined as follows: HH (number of hairs more than 55), hh (hairless septa), or Hh (number of hairs from 1 to 55) genotype. Durrant and Joarder (1978) also concluded that, if more than 55 hairs are present on the seed capsule septa, then most likely, but not necessarily, this plant may have the LIS-1 insertion. The heterozygous condition Hh was shown to be unstable due to gene interaction and could be visually detected by the number of hairs on the septa, which varied from about 20 to 55 (Durrant and Joarder, 1978).

We counted the number of hairs on the capsule septa for seven flax varieties: 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), 624_6219 (R2 group),

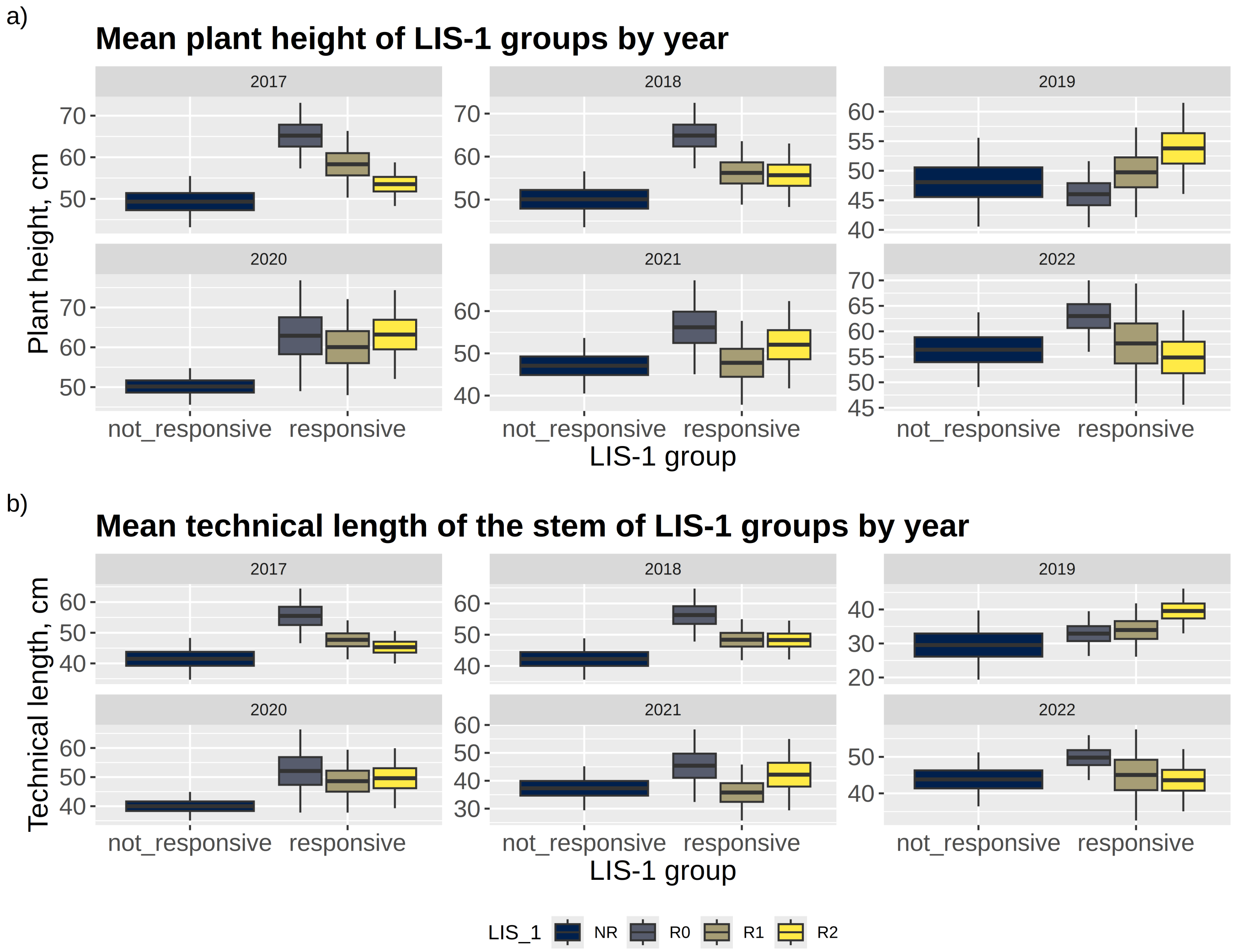

Figure 3. Comparison of quantitative morphological characteristics by four LIS-1 groups over six years where ANOVA showed significant results. a) mean height of the plant (cm); b) mean technical length of the stem (cm) for R0, R1, R2 groups ('responsive group'), and NR ('non-responsive').

and 624_791 (R0 group). The number of hairs varied from 0 to about 65 (Figure 4). According to Durrant and Joarder (1978), we assigned plants with hairless septa as recessive homozygotes (hh), plants with more than 55 hairs as dominant homozygotes (HH) and others as heterozygotes (Hh). In Figure 4, the different types of ciliation of the septa are shown.

In our study, dominant homozygotes HH were only present in varieties from the LIS-1 group R1 (624_6222, 624_1044), which have the LIS-1 insertion, whereas recessive homozygotes hh were represented in all LIS-1 groups, except varieties 624_6222 and 624_1044, that were dominant homozygotes HH or heterozygotes Hh, respectively (Figure 5).

Prediction of LIS-1 presence using the machine learning algorithm random forest classifier

Using the machine learning algorithm random forest classifier (Breiman, 2001), we tried to predict the presence, absence or heterozygosity of the LIS-1 insertion based on the characteristics of plants. This classifier also allowed us to calculate the importance of each characteristic of the object for the classification. The features used in the machine learning model were: ciliation of the seed capsule's septa (hairless hh, heterozygous Hh, hairy HH), plant height, technical length of the stem (as they are shown in literature to be associated with LIS-1), number of productive seed capsules per plant, and number of seeds in the capsule (based on correlation analysis), presence (weak or missing) of anthocyanin pigmentation in the hypocotyl.

The target variable was the presence of LIS-1 insertion ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote'). The heterozygosity of LIS-1 could be attributed to the insertion occurring in one chromosome, while being absent in the other. This condition is unstable, and due to uncertainty regarding its transmission in the next generation, we encoded it separately.

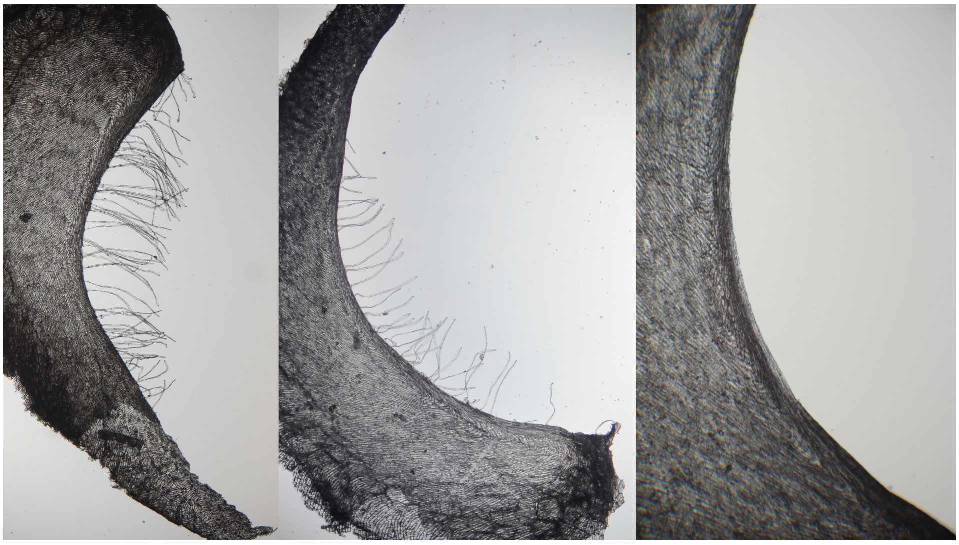

Figure 4. Ciliation of septa of flax genotypes. From left to right: dominant homozygous HH (624_6222), heterozygous Hh (624_6215) and recessive homozygous hh (624_791).

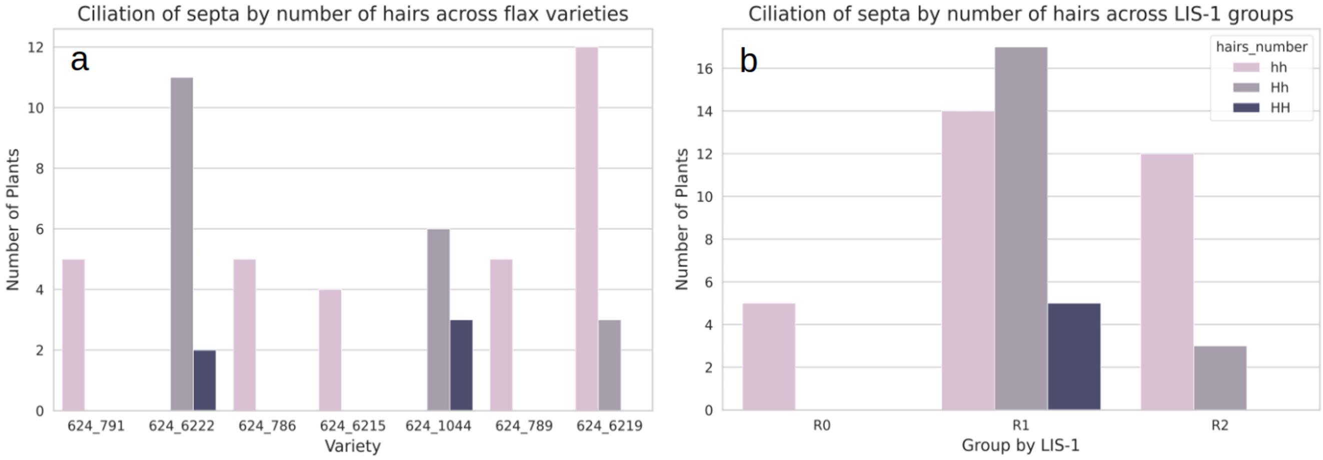

Figure 5. Distribution of hairy septa feature of plants across flax varieties (a) and LIS-1 groups (b). Accessions 624_791 (R0 group), 624_6222, 624_1044, 624_786, 624_789, 624_6215 (R1 group), and 624_6219 (R2 group); groups by LIS-1: 0 – R0, 1 – R1, 2 – R2, N = 5, 37, 15 plants.

Overall, the classification accuracy of the predictive model was 99.14% on training data and 98.039% on test data. Of the 51 objects, 34 were correctly classified as 'LIS-1 present', 9 belonged to 'LIS-1 absent', and 7 were 'heterozygotes LIS-1'. One object was erroneously assigned to 'LIS-1 heterozygotes' (in fact, 'LIS-1 present', Figure 6). So, groups with 'LIS-1 present' and 'LIS-1 heterozygote' based on studied morphological characteristics could be misclassified.

Classification metrics that show the success of the class prediction were not equal for all studied groups of flax varieties. The average accuracy of predictions was 98%. The general classification report is given in Table 9.

The random forest classifier calculates the importance of each characteristic of the object (shown in Figure 7) for the classification task.

Five morphological features had classification importance exceeding 10%: technical stem length (19.86%), plant height (19.33%), number of capsules (18.73%), ciliation of septa hh (means hairless septa, 11.68%), and number of seeds per capsule (10.61%).

Discussion

The ancient local flax varieties investigated in this study originated from different regions of Belarus. They are adapted to certain growing conditions in which they have evolved and could therefore serve as useful sources of genetic diversity. Surveying and inventorying the pool of diversity in local varieties, including using molecular markers, is a priority to sustain future agricultural production (FAO, 2012).

Genome plasticity could be defined as a change in genome structure (mutations, genome expansion, transposable elements, etc.) associated with environmental challenges, leading to the development of new phenotypes. Meanwhile, adaptive plasticity is a phenotypic

Table 9. Classification metrics for each group of plants.

	Precision	Recall	F1-score	Number of objects
LIS-1 absent	100	100	100	9
LIS-1 present	100	97	99	35
LIS-1 heterozygotes	88	100	93	7
Accuracy			**98**	

plasticity that increases the global fitness of a genotype. Genotype fitness refers to the relative abundance and success of a species' genes over multiple generations (total biomass, seed number and growth rates of a single generation) (Nicotra *et al*, 2010).

The majority of studied flax varieties were assigned to the responsive group of genotypes, defined by presence of the LIS-1 insertion (groups R0, R1 and R2, which included 18 of the studied accessions), and only three accessions (assigned to NR group) were not responsive in terms of LIS-1.

For plant height and technical length of the stem, the groups R0 (that formed the insertion and then lost it) and R2 (that formed the insertion, and partly lost it) were statistically significant ($p < 0.05$) higher than group NR (the insertion was not found). Group R1 included the accessions that formed the LIS-1 insertion and retained it. Statistically, there was no significant difference observed between group R1 and groups R0 and R2 ($p > 0.05$) by plant height and technical length of the stem and at the same time, group R1 was not significantly higher than the NR group ($p > 0.05$). Bickel *et al* (2012) indicated that stable S- and L-genotrophs (which have retained and lost the insertion sequence) were well adapted to environmental stress (lack and excess nitrogen or water in the soil, respectively) compared to the PL line, which in normal conditions does not form the insertion, but under stressful conditions could produce two types of stable S- and L-genotrophs, as well as retain the ability to be genetically plastic. Flax varieties that stably inherit the LIS-1 insertion (S-genotrophs) are characterized by a shorter plant height than L-genotrophs and the PL-line (Bickel *et al*, 2012). This could be attributed to the LIS-1 insertion being one of multiple genomic rearrangements occupying a region with two genes involved in growth processes, inhibitor of growth-1, and kip-related cyclin-dependent kinase inhibitor-2 (Bickel *et al*, 2012). Thus, it could affect both the plant height and technical stem length. The weather conditions in the years 2017 to 2022 differed in terms of rainfall and temperatures during the flax vegetation period. The lack of a statistically significant interaction effect between genotype (R1 group) and June temperatures on plant height ($p = 0.108$) indicates that these genotypes may be phenotypically stable, and would not modify their height in response to temperatures in the stage of active growth. An effect on the border of significance was revealed for genotype (NR group) × Rainfall_May interaction for the reproduction-related trait 'number of seeds in the capsule' ($p= 0.056$). This indicated that while this group of accessions did not exhibit genetic responsiveness, they displayed phenotypic plasticity. Significant negative correlations between growth-related traits (plant height and technical stem length) and LIS-1 groups showed that the absence of the insertion (NR group) is associated with shorter plant height and stem length.

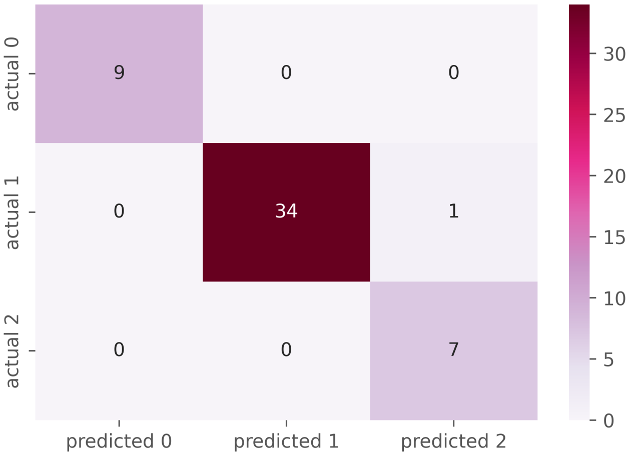

Figure 6. Confusion matrix showing classification accuracy of prediction of LIS-1 presence by machine learning. Of the 51 objects, 50 are correctly classified. One object was misclassified. ('0' – 'LIS-1 absent', '1' – 'LIS-1 present', '2' – 'LIS-1 heterozygote').

For reproduction-related traits ('number of seed capsules per plant' and 'number of seeds in the capsule'), responsive genotypes (R0, R1 and R2) did not significantly differ from non-responsive genotypes (NR, $p > 0.05$).

Flax is mostly inbred and, therefore, under selection by the environment for particular combinations of alleles, it could become homozygous at all loci with little variation (Cullis, 2019), making it vulnerable to any environmental changes. Therefore, the ability to modify the genome in response to growth challenges could be an evolutionary advantage. The uniqueness of the genome plasticity mechanism in fibre flax is that modification occurs not in a single gene but in different genome regions, resulting in potential phenotypic and biochemical variability (Cullis, 2019). Thus, the LIS-1 sequence is a promising molecular marker for identifying flax forms with genome plasticity.

A complex of phenotypic changes in the stable L- and S-genotrophs are described in the literature (Bickel *et al*, 2012; Cullis, 2019). Among them are the height of the plant, the hairy septa, and the number of

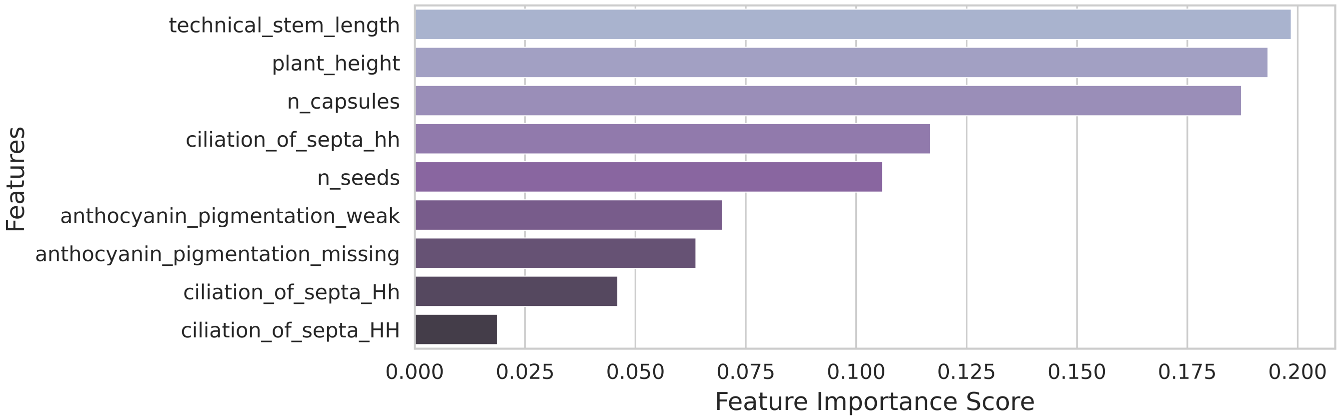

Figure 7. Feature importance score calculated by random forest classifier. The features were: technical length of the stem, plant height, number of seed capsules per plant (n_capsules), ciliation of the seed capsule's septa (hairless hh, heterozygous Hh, hairy HH), and number of seeds in the capsule (n_seeds), anthocyanin pigmentation in the hypocotyl (weak or missing).

capsules, which were the most important for our random forest classification model. The overall accuracy of LIS-1 status (presence, absence or heterozygous condition) prediction based on nine studied morphological features was 98.039%.

From the evolutionary point of view, both LIS-1 responsive and non-responsive flax varieties are successful in transmitting their genes over generations as they do not differ by the number of seed capsules and seeds. The study of plant genetic resources using LIS-1 as a molecular marker of genome plasticity will provide us with knowledge about flax genotypes that are potentially valuable for fibre flax breeding and biodiversity conservation.

Conclusion

Among the 21 local varieties studied, four groups were identified based on their ability to modify their genome using LIS-1 as a molecular marker of genome plasticity. The most promising in terms of sources for the selection of fibre flax varieties adaptive to environmental challenges is the group of responsive varieties that have formed LIS-1 insertion (R0, R1 and R2 groups). Existing associations between patterns of LIS-1 sequence presence and morphological traits of flax allowed us to classify them correctly with 98% accuracy.

Supplemental data

Supplemental Table 1. Characteristics of flax varieties for machine learning modelling

Acknowledgements

We thank Viachaslau N. Kipen for providing photos of the hairy and hairless septa of flax capsules.

The work was carried out within the framework of the activity 'Genetically identify collection samples of agricultural crops to form a new gene pool of donors of economically valuable traits for use in breeding', subprogramme 'Study, identification and rational use of collections of plant genetic resources' of the state programme 'Scientific and innovative activities of the National Academy of Sciences of Belarus' for 2021–2025.

Conflict of interest statement

The authors have no conflicts of interest to report.

Author contributions

MP contributed to the study conception, writing of the draft and final manuscript, statistical analysis, modelling, visualization and interpretation of the results, laboratory experiments conduction, and final manuscript revision. VL contributed to the study conception and design, interpretation of the results, final manuscript revision, resource provision and supervision of the research on field and laboratory experiments. EL contributed to the study conception and design, laboratory experiments conduction, writing of the final manuscript and interpretation of the results. VS contributed to the study conception and design, field experiment design and conduction, made field measurements, and worked on the draft. AB contributed to field experiment design and conduction, and worked on the draft, study conception and design. EG contributed to the laboratory experiments design and conduction, worked on the draft, and to the study conception and design. LK contributed to the study conception and design, and the final manuscript revision. All authors discussed the results and commented on the manuscript, and have read and agreed to the published version of the manuscript.

References

Bajorath, J. (2022). Revisiting active learning in drug discovery through open science. *Artificial Intelligence in the Life Sciences* 2, 100051–100051. doi: https://doi.org/10.2144/fsoa-2022-0010

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. *Jour-*

nal of Statistical Software (1), 67–67. doi: https://doi.org/10.18637/jss.v067.i01

Bickel, C., Lukacs, and Cullis, C. (2012). The loci controlling plasticity in flax. *Research and Reports in Biology* 3, 1–11. doi: https://doi.org/10.2147/RRB.S27198

Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. doi: https://doi.org/10.1023/A:1010933404324

Chen, Y., Lowenfeld, R., and Cullis, C. A. (2009). An environmentally induced adaptive (?) insertion event in flax. *Journal of Genetics and Molecular Biology* 1, 38–047.

Chen, Y., Schneeberger, R. G., and Cullis, C. A. (2005). A site-specific insertion sequence in flax genotrophs induced by environment. *The New Phytologist* 167, 171–80. doi: https://doi.org/10.1111/j.1469-8137.2005.01398.x

Cullis, C. (1976). Environmentally induced changes in ribosomal RNA cistron number in flax. *Heredity* 36, 73–79. doi: https://doi.org/10.1038/hdy.1976.8

Cullis, C. A. (1981). DNA sequence organization in the flax genome. *Biochim Biophys Acta* 652(1), 1–15. doi: https://doi.org/10.1016/0005-2787(81)90203-3

Cullis, C. A. (1986). Phenotypic consequences of environmentally induced changes in plant DNA. *Trends in Genetics* 2(86), 90285–90289. doi: https://doi.org/10.1016/0168-9525(86)90285-4

Cullis, C. A. (2019). Origin and Induction of the Flax Genotrophs. *Genetics and Genomics of Linum* 227-234. doi: https://doi.org/10.1007/978-3-030-23964-0_14

Diederichsen, A. (2019). A Taxonomic View on Genetic Resources in the Genus *Linum* L. for Flax Breeding. *Genetics and Genomics of Linum* 227-234. doi: https://doi.org/10.1007/978-3-030-23964-0_1

Durrant, A. and Joarder, O. I. (1978). Regulation of hairless septa in flax genotrophs. *Genetica* 48, 171–183. doi: https://doi.org/10.1007/BF00155567

Durrant, A. and Jones, T. (1971). Reversion of induced changes in amount of nuclear DNA in *Linum*. *Heredity* 27, 431–439. doi: https://doi.org/10.1038/hdy.1971.106

Durrant, A. and Nicholas, D. (1970). An unstable gene in flax. *Heredity* 25, 513–527. doi: https://doi.org/10.1038/hdy.1970.60

Ehrensing, D. (2008). Oilseed Crops: Flax (EM 8952-E) (Oregon State University Extension Service).

Evans, G., Durrant, A., and Rees, H. (1966). Associated Nuclear Changes in the Induction of Flax Genotrophs. *Nature* 212, 697–699. doi: https://doi.org/10.1038/212697a0

FAO (2012). Synthetic account of the Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture. url: https://www.fao.org/3/i2650e/i2650e.pdf.

Goldsbrough, P. B., Ellis, T. H., and Cullis, C. A. (1981). Organisation of the 5S RNA genes in flax. *Nucleic Acids Res* 9(22), 5895–904. doi: https://doi.org/10.1093/nar/9.22.5895

Harris, C. R., Millman, K. J., and Van Der Walt, S. J. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: https://doi.org/10.1038/s41586-020-2649-2

Hu, Z. and Xing, E. P. (2021). Toward a 'Standard Model' of Machine Learning. *Harvard Data Science Review* . doi: https://doi.org/10.1162/99608f92.1d34757b

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9(3), 90–95. doi: https://doi.org/10.1109/MCSE.2007.55

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. url: https://rpkgs.datanovia.com/rstatix/.

Kluyver, T. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas,* ed. Loizides, F. and Schmidt, B. 87-90.

Maggioni, L., Pavelek, M., Van Soest, L. J. M., and Lipman, E. (2001). Flax Genetic Resources in Europe Ad hoc meeting (Rome, Italy: International Plant Genetic Resources Institute), 72-73. url: https://www.ecpgr.cgiar.org/fileadmin/bioversity/publications/pdfs/Ad_Hoc_Fibre_Crops_WG_ad_hoc_meeting_Flax_genetic_resources_in_Europe_Czech_Rep_2001.pdf.

Nicotra, A. B., Atkin, O. K., Bonser, S. P., Davidson, A. M., Finnegan, E. J., Mathesius, U., Poot, P., Purugganan, M. D., Richards, C. L., Valladares, F., and Van Kleunen, M. (2010). Plant phenotypic plasticity in a changing climate. *Trends in plant science* 15(12), 684–692. doi: https://doi.org/10.1016/j.tplants.2010.09.008

Nôžková, J., Pavelek, M., Bjelková, M., Brutch, N., Tejklová, E., Porokhovinova, E., and Brindza, J. (2016). Descriptor list for flax (*Linum usitatissimum* L.) . doi: https://doi.org/10.15414/2016.9788055214849

Oliveros, J. C. (2007-2015). Venny. An interactive tool for comparing lists with Venn's diagrams. url: https://bioinfogp.cnb.csic.es/tools/venny/index.html.

Privalov, F. I., Grib, S. I., and Matys, I. S. (2021). National seed bank of genetic economically useful plant resources is a scientific object of a National property of the Republic of Belarus. *Crop Farming and Plant Growing* 2, 10–14.

Rachinskaya, O. A., Lemesh, V. A., Muravenko, O. V., Yurkevich, O., Yu, Guzenko, E. V., Bol'sheva, N. L., Bogdanova, M. V., Samatadze, T. E., Popov, K. V., Malyshev, S. V., Shostak, N. G., Heller, K., Hotyleva, L. V., and Zelenin, A. V. (2011). Genetic polymorphism of flax *Linum usitatissimum* based on the use of molecular cytogenetic markers. *Russ J Genet* 47, 56–65. doi: https://link.springer.com/article/10.1134/S1022795411010108

Raghunathan, S. and Priyakumar, U. D. (2022). Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chem-*

istry 122(7), 26870–26870. doi: https://doi.org/10.1002/qua.26870

Sa, R., Yi, L., Siqin, B., An, M., Bao, H., Song, X., Wang, S., Li, Z., Zhang, Z., Hazaisi, H., Guo, J., Su, S., Li, J., Zhao, X., and Lu, Z. (2021). Chromosome-Level Genome Assembly and Annotation of the Fiber Flax (*Linum usitatissimum*) Genome. *Front Genet* 12, 735690–735690. doi: https://doi.org/10.3389/fgene.2021.735690

Sambrook, J. and Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol:chloroform. *CSH protocols* . doi: https://doi.org/10.1101/pdb.prot4455

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference.

The pandas development team (2020). pandas-dev/pandas: Pandas 1.0.0 (v1.0.0). Zenodo. url: https://doi.org/10.5281/zenodo.3630805.

Vavilov, N. I. (1926). Studies on the origin of cultivated plants. *Bull Appl Botany* 16(2), 3–248.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., Vanderplas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and Van Mulbregt (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17(3), 261–272. doi: https://doi.org/10.1038/s41592-019-0686-2

Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine Learning for Small Molecule Drug Discovery in Academia and Industry. *Artificial Intelligence in the Life Sciences* . doi: https://doi.org/10.1016/j.ailsci.2022.100056

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software* 6(60), 3021–3021. doi: https://doi.org/10.21105/joss.03021

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis (Springer International Publishing), 2nd edition. url: https://ggplot2.tidyverse.org.

Wickham, H., Franc üller, K. (2022). dplyr: A Grammar of Data Manipulation. url: https://dplyr.tidyverse.org.

Yang, Z., Tian, Y., Kong, Y., Zhu, Y., and Yan, A. (2022). Classification of JAK1 Inhibitors and SAR Research by Machine Learning Methods. *Artificial Intelligence in the Life Sciences* 2, 100039–100039. doi: https://doi.org/10.1016/j.ailsci.2022.100039