Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Mapping pea seed composition through strategic selection of accessions from the Nordic gene bank☆

Qinhui Xing[a], Zhi Ye[b,c], Bo Yuan[a], Xiaoxiao Liu[a], Morten Arendt Rasmussen[b,c], Jacob Judas Kain Kirkensgaard[d,e], Michael Lyngkjær[f], Ulrika Carlson-Nilsson[f], Cecilia Hammenhag[g], Rene Lametsch[a,*]

[a] *Food Analytics and Biotechnology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark*
[b] *Food Microbiology, Gut Health, and Fermentation, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark*
[c] *Copenhagen Prospective Studies on Asthma in Childhood, Herlev-Gentofte Hospital, Copenhagen University Hospital, Denmark*
[d] *Ingredient and Dairy Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark*
[e] *Condensed Matter Physics, Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 København Ø, Denmark*
[f] *NordGen Plants, Nordic Genetic Resource Center, Växthusvägen 12, Alnarp, Sweden*
[g] *Department of Plant Breeding, The Swedish University of Agricultural Sciences, Sundsvägen 10, Alnarp, Sweden*

ARTICLE INFO

Keywords:
Pea accessions
Algorithm-assisted selection
Image features
Compositions
Mapping
Plant-based ingredients

ABSTRACT

This study aims to utilise natural variation in pea seed composition from NordGen collections to identify key traits for optimized plant-based ingredients functionality while minimizing refined extraction processes. Given the impracticality of chemically analysing 1942 accessions, an algorithm-assisted approach was employed, using image-derived features and datasets to pre-select 51 accessions. Protein content, thousand kernel weight, perimeter, and G-value were determined as primary criteria via PCA, capturing variations in protein composition and other key components. Protein and starch content ranged from 21.2 to 36.9 % and 21.0–48.1 %, respectively. Image analysis linked geometry to composition, aiding pea selection and application. X-ray scattering differentiates peas based on starch structure. Proteomic profiling revealed that legumin and vicilin varied most, with legumin dominant in smooth peas and vicilin in wrinkled ones, enabling control of their ratio through selection. This study highlights the potential of using natural variation of seed composition for less-refined plant-based ingredients for various applications.

1. Introduction

As global demand for plant-based foods continues to rise, peas have emerged as a valuable protein source due to their high yield and relatively low environmental footprint (Saget et al., 2021). A major challenge in using pea ingredients for meat analogues and other plant-based applications is their complex composition of protein, starch, and fiber. This complexity, along with their undesirable sensory attributes, can result in poor food texture and reduced consumer acceptance (C. Sun et al., 2021). A common approach to overcoming these challenges is to use advanced fractionation technologies to separate key compounds, such as proteins and starch, from a small number of high-yielding pea cultivars (Kornet et al., 2022; Lie-Piang et al., 2025; Pelgrom et al., 2013). The selected fractions are then used in food formulation. While this method improves functionality, it relies on a limited set of cultivars that restricts the diversity of sensory and functional traits needed for diverse plant-based applications, leaving much of the potential among non-commercial cultivars unexploited. As a result, many plant-based products require specific attributes that are often lacking in modern, highly refined cultivars. Besides, the sustainability of fractionation is also questionable since it usually requires significant energy and water inputs, raising concerns about its high environmental cost. These challenges underscore the need for alternative strategies that leverage genetic diversity to enhance the functional properties of pea proteins while reducing reliance on resource-intensive processing methods.

The Nordic Genetic Resource Center – NordGen (*WebSource: Plants –*

☆ This article is part of a Special issue entitled: '3rd NIZO Plant Protein' published in Food Chemistry.

* Corresponding author.
E-mail address: rla@food.ku.dk (R. Lametsch).

https://doi.org/10.1016/j.foodchem.2025.145478
Received 24 March 2025; Received in revised form 30 June 2025; Accepted 7 July 2025
Available online 11 July 2025
0308-8146/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

NordGen) maintains a collection of over 2700 pea accessions, offering a rich source of genetic diversity. These accessions exhibit natural variations in key components such as carbohydrate and protein compositions, which play a crucial role in determining processing properties. For instance, mutations in the *Rug3* gene result in starch-free peas, while mutations at the *Rb* loci and *Vc-2* locus significantly reduce the levels of legumin and vicilin, respectively (B. Chen et al., 2024). This genetic diversity provides opportunities to identify accessions with superior functional and nutritional properties tailored to specific food applications. However, despite this diversity and the historical significance of pea as a genetic model, its genomics research has lagged behind other major legumes due to the large size and complexity of its genome (Pandey et al., 2021). As a result, systematic phenotypic characterisation provides a practical alternative to guide raw material selection and trait discovery for plant-based food development.

Previous studies have explored compositional and functional differences across selected accessions, for instance, comparative proteomic analyses across eight cultivars to reveal differences in protein composition and genetic variants, while functional and sensory profiling of isolates from twelve pea cultivars highlighted variability relevant to food applications (Arteaga et al., 2021; Vreeke et al., 2023). However, large-scale and integrative analyses remain underutilised. The lack of genotypic data (e.g., SNPs) for many accessions restricts the ability to link natural variation with underlying genetic factors, hindering the application of genome-informed selection strategies. Despite this potential, the practical utilisation of genebank material presents several challenges. One major limitation is the small quantity of seeds available for each accession, which restricts the extent of destructive testing required for compositional and functional analysis. This constraint makes it difficult to assess large numbers of accessions comprehensively, limiting the ability to identify superior traits for food applications. Additionally, genebank collections often lack detailed processing-related data, necessitating resource-intensive screening methods to uncover functional properties relevant to modern plant-based foods.

Detailed compositional and functional analysis of genebank accessions is crucial for identifying superior traits that can enhance plant-based food applications (Nguyen & Norton, 2020). Advanced methods, such as non-destructive imaging and high-throughput compositional assays, are essential for maximizing information from limited seed quantities. This approach can potentially be used to select seeds with large natural variation in the seed composition and accelerate the identification of accessions with specific composition and superior functional properties. Ultimately, effective characterisation and targeted breeding can transform the use of peas in modern food systems, leading to next-generation cultivars tailored for plant-based applications.

To address these challenges, this study aims to apply a high-throughput image processing approach with algorithm-assisted clustering models to predict and assess pea seed compositions. This approach enables the identification of a representative subset of accessions capturing a wide range of compositional variations. These selected accessions can then be used to study the functional impact of individual compounds in greater detail. Overall, this data-driven strategy offers a rapid and cost-effective alternative to traditional wetlab analyses, facilitating the efficient screening of large pea germplasm collections for the development of functional plant-based food ingredients.

2. Materials and methods

2.1. Clustering-based selection of genebank accessions

A database containing 53 descriptors of 1942 genebank pea accessions, along with images, was provided by NordGen. Detailed information can be found on the Nordic Baltic Genebanks Information System (GenbIS, Search Accessions GRIN-Global). From this dataset, a highly diverse panel of pea accessions was selected for further analysis.

A subset of 51 accessions was selected from the NordGen pea collection based on four standardized variables: protein content, thousand kernel weight (TKW), perimeter, and G-value. Protein content and TKW were extracted from the GenbIS, while perimeter and G-value were derived from image analysis (see Section 2.3). These variables were selected based on their high explanatory power in correlation and principal component analysis (PCA) (explained in Section 3.2). The Partitioning Around Medoids (PAM) clustering algorithm (k = 50) was applied to the dataset (n = 1448) to identify representative medoid samples across the diversity space. Extreme values were also retained to ensure edge coverage. A total of 51 accessions were included for further study after confirming seed availability. To ensure conservation, they were stored at −18 °C in hermetically sealed aluminum bags with a moisture content below 7 %. The seeds were stored at −18 °C in hermetically sealed aluminum bags with a moisture content below 7 % to ensure conservation.

2.2. Raw materials

Whole pea seeds were milled into fine flour using a mixer mill MM 400 (Retsch, Verder Scientific GmbH & Co. KG, Germany) equipped with two 10 mL agate grinding jars. For each batch of milling, two to three pea seeds (∼2 mL) and two agate beads with a diameter of 10 mm were loaded into the jar and fixed on the vibration arm. The seeds were vibrated vigorously at a maximum frequency of 30 Hz (1800 min^{-1}) for 2 min. Grinding was performed at room temperature. The resulting flour was sieved through a 0.25 mm mesh, and the grinding jars were rinsed with water between samples. The fine flour was collected into a 4 mL polypropylene bottle with a threaded screw cap and stored at −20 °C before use.

2.3. Image processing for feature extraction

2.3.1. Mask generation using segment anything model

The images of peas used in this study were sourced from NordGen Plants and consist of high-quality visuals of various pea accessions captured under standardized conditions using a Tagarno video microscope equipped with a 5.0× lens (Fig. 1. A). The photographs were taken against a uniform white background, with the camera positioned vertically, perpendicular to the ground, and at a fixed distance from the peas. All images in this study had the same resolution of 1920 × 1080 pixels, with RGB channels. This setup ensured consistency across the dataset, making it suitable for accurate segmentation. Each image represents a pea accession, with a total of 2030 images corresponding to 1942 accessions, as some accessions have multiple images.

The Segment Anything Model (SAM) was used for automatic mask generation (Fig. 1. B). SAM is a highly generalized image segmentation model developed by Meta AI (Kirillov et al., 2023). It was pre-trained using 1.1 billion high-quality segmentation masks derived from 11 million diverse images to enable promptable segmentation tasks. SAM's design was inspired by Natural Language Processing (NLP) models, where a foundation model is pre-trained to predict the next token and can be adapted to downstream tasks via zero-shot transfer. Similarly, SAM is assumed to possess a generalized understanding of objects, enabling it to effectively segment them from images (Mazurowski et al., 2023). SAM consists of two encoders and one decoder. The model uses a Masked Autoencoder (MAE) (He et al., 2022) to generate the image embedding, along with a prompt encoder and a mask decoder to map all embeddings to corresponding segmentation masks. Additionally, beyond using direct prompts, SAM can generate masks automatically without prompts. It samples a large number of prompts across an image to generate masks for all objects present.

The images of peas contained a single object category. Masks were generated for all peas in the image after quality filtering, with each mask covering a single pea. The box IoU (Intersection over Union) cutoff was lowered to prevent small masks from overlapping with the primary

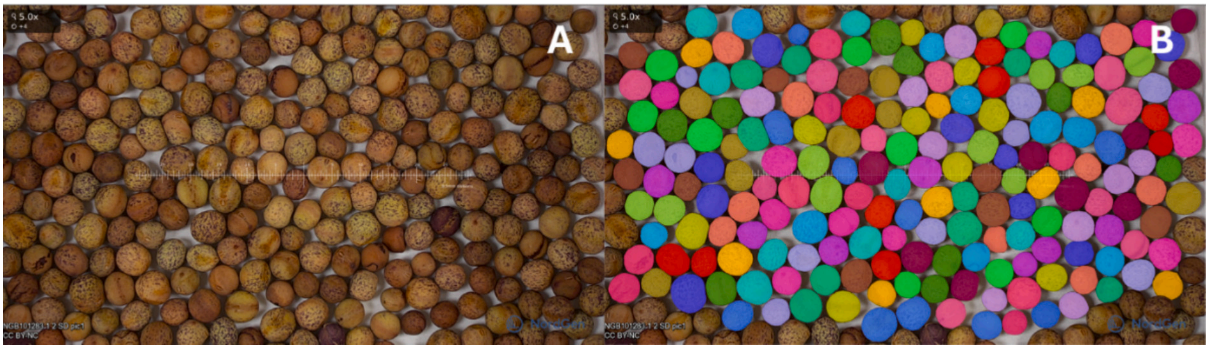

Fig. 1. Schematic diagram of using SAM to process pea images. A: original pea image, B: processed pea image, each single-coloured enclosed shape is referred to as a mask.

objects of interest, and a minimum mask size was set to exclude insignificant masks. A total of 2030 images were input into the SAM to generate pea masks. Only masks fully contained within the image were included, excluding those whose bounding boxes extended beyond the image boundaries. This step ensured accurate computation of area- or shape-related features of the peas. Despite this adjustment, the remaining masks still captured the overall characteristics of each accession due to the high number of peas present in the images and their central positioning. Finally, a total of 286,862 masks were identified from 1942 accessions.

2.3.2. Features computation

For each mask, features including area, perimeter, major axis length, minor axis length, eccentricity, solidity, roundness, and RGB channel values were computed. Among them, area, perimeter, major and minor axis lengths are geometric features that describe the size of seeds. Eccentricity, solidity, and roundness are parameters used to describe the shape of seeds. The mean and standard deviation of these features were then calculated to represent the overall image features for each accession. All feature extraction steps were performed using scikit-image (v.0.19.2) (Van Der Walt et al., 2014). Automatic mask generation was performed using PyTorch (v.1.13.0) (Paszke et al., 2019), and the pipeline was implemented in Python (v.3.10.8).

Area: The number of pixels in the mask.

Perimeter: The total length of the line approximates the mask's contour.

Major axis length: The length of the major axis of the mask, calculated by fitting an ellipse with the same normalized second central moments as the mask.

Minor axis length: The length of the minor axis of the mask.

Eccentricity: The ratio of the distance between focal points to the major axis length ranges between 0 and 1. This ratio equals 0 for a perfect circle; otherwise, the shape resembles an ellipse.

$$\text{Eccentricity} = {}^{\text{Distance between focal points}}/_{\text{Major axis length}}$$

Solidity: The ratio of the mask's area to the area of its convex hull. Solidity equals 1 when the shape has no indentations.

$$\text{Solidity} = {}^{\text{Mask's area}}/_{\text{Convex hull area}}$$

Roundness: The degree of similarity between the mask's shape and that of a mathematically perfect circle. Roundness equals 1 for a perfect circle and decreases as the shape becomes more polygonal.

$$\text{Roundness} = {}^{4\pi \times \text{Area}}/_{\text{Perimeter}^2}$$

RGB values: The median and standard deviation of pixel values in the red, green, and blue channels within the mask.

2.4. Selection of representative accessions

The dataset of 1448 accessions was clustered into 50 groups using the Partitioning Around Medoids (PAM) algorithm to select a representative subset of pea accessions for further analysis. Four key variables (protein content, KTW, perimeter, and G-value) were normalized, and clustering was performed with $k = 50$. The medoid sample of each cluster was chosen as a representative. To ensure comprehensive trait coverage, accessions with extreme protein content values (i.e., maximum and minimum) were included if not already represented. After confirming seed availability with NordGen, a total of 51 accessions were selected. The panel consisted of 51 accessions of *Pisum sativum* L., and one *Pisum abyssinicum* A. Braun accession. Among these, 20 accessions originated from Sweden, 7 from Russia, 4 each from Germany and Bulgaria. Two accessions were from the Netherlands, and one each from Denmark, Finland, Czechia, Ethiopia, Afghanistan, and Bhutan. The origins of the remaining 8 cultivars were undocumented.

2.5. Compositional analysis

2.5.1. Protein content

Approximately 20 mg of pea flour was weighed onto a 35 × 35 mm tin foil (Elementar Analysensysteme GmbH, Langenselbold, Germany) using an analytical balance (AG 135, METTLER TOLEDO, Switzerland). The tin foil was then folded into a capsule and loaded into the autosampler. The protein content was determined by the Dumas combustion method with an organic elemental analyser (vario MACRO cube, Elementar Analysensysteme GmbH, Germany). The operational parameters were as follows: combustion tube temperature, 960 °C; reduction tube temperature, 830 °C; pressure, 1230 mbar; helium flow rate, 600 mL/min; MFC-TCD, 600 mL/min. A nitrogen conversion factor of 5.4 was used to calculate crude protein content (Vreeke et al., 2023). The measurements were performed in duplicate, and results are reported on a dry weight basis.

2.5.2. Total starch

The starch content of all samples was determined using the Megazyme Total Starch Assay Kit (Wicklow, Ireland), following the manufacturer's instructions.

2.5.3. Amylose/amylopectin ratio

The ratio of amylose to amylopectin in pea flour was determined using the commercial Megazyme Amylose/Amylopectin Assay Kit with some modifications. Specifically, 25 mg of pea flour was mixed with 0.5 mL of 80 % (*v*/v) ethanol in a 15 mL Falcon tube and vortexed thoroughly. Then, 6 mL of 96 % (v/v) ethanol was added, and the mixture was allowed to stand for 15 min. The tube was centrifuged at 5000 ×g for 5 min, and the supernatant was discarded. The pellet was drained on tissue for 15 min. Subsequently, 0.1 mL of 80 % ethanol was added, and the sample was vortexed to disperse the pellet. Afterward, 1 mL of cold 1.7 M sodium hydroxide was added, and the mixture was vortexed for 15 s. The tubes were then placed on ice for 15 min. Next, 4 mL of 0.6 M sodium acetate buffer (pH 3.8) containing 5 mM calcium chloride was added. The total volume was brought to 25 mL before proceeding with the subsequent steps as described in the kit protocol. The amylose content was expressed as a percentage on a dry starch basis.

2.6. *Wide-angle X-ray scattering (WAXS)*

Pea seeds were dehulled and milled into fine flour using a mixer mill. The samples were sealed in a 0.5 mm thick mica cell and scanned from $2\theta = 5^\circ$ to 40° using a Nano-inXider instrument (Xenocs, Grenoble, France). X-rays were generated from a Cu Kα source at settings of 40 mA current and 40 kV voltage, with a wavelength of 1.54 Å. Scattering from an empty sample cell was subtracted as background. Each sample was measured three times at different positions, and the average spectrum was used for analysis. The q range of 0.41–4.20 Å^{-1} (where $q = 4\pi \sin \theta/\lambda$) was utilised. Crystallinity (%) was calculated as the ratio between the area of the crystalline peaks and the amorphous baseline using MATLAB (version 2022a, The MathWorks, Inc., Massachusetts, USA). The baseline was estimated using a robust smoothing algorithm proposed by Brückner (2000).

2.7. *Subunit composition of pea flour analysed by LC-MS/MS*

2.7.1. *In-solution trypsin digestion*

Pea flour (50 mg) was solubilized in 5 mL of 100 mM Tris-HCl buffer (pH 8.0, containing 8 M urea) and thoroughly mixed overnight at room temperature on a rotator (MX-RL-E, DLAB Scientific Co., Ltd., Beijing, China). The protein concentration of the solution was adjusted to 1.0 mg/mL using 100 mM Tris-HCl buffer. In-solution trypsin digestion was performed following the method described by Zhang et al. (2024) with minor modifications. A 20 μL aliquot of the protein solution was transferred to a 1.5 mL Eppendorf tube, followed by the addition of 4 μL of 450 mM dithiothreitol (DTT). The mixture was incubated at room temperature for 45 min to reduce disulphide bonds. Subsequently, 8 μL of freshly prepared 500 mM iodoacetamide (in 100 mM NH_4HCO_3) was added to alkylated native and reduced cysteine residues, and the sample was incubated in the dark for 1 h at room temperature. Then, 158 μL of 10 mM NH_4HCO_3 was added to adjust the protein concentration to 0.1 μg/μL. Proteolysis was initiated by adding 10 μL of 0.125 μg/μL trypsin (from bovine pancreas, activity ≥10,000 BAEE, Sigma-Aldrich) and incubating the mixture overnight at 37 °C. The hydrolysis was terminated by adding 4 μL of 10 % (*v*/*v*) trifluoroacetic acid, giving a pH of 2.0–2.5. The digested sample was centrifuged at 20,000 ×*g* for 20 min, and the supernatant was filtered through a 0.2 μm regenerated cellulose (RC) membrane filter. An aliquot of 50 μL of the filtrate was transferred to a 96-well PCR plate and stored at −20 °C prior to LC-MS analysis.

2.7.2. *LC-MS/MS*

The tryptic peptides were analysed using an Orbitrap Exploris 480 mass spectrometer coupled with a Vanquish UHPLC system (Thermo Fisher, Roskilde, Denmark). The peptides were separated by a bioZen™ Peptide XB-C18 column (1.7 μm particle size, 150 × 2.1 mm, Phenomenex, Værløse, Denmark) and eluted in 40 min using a linear gradient of solvent A (0.1 % formic acid) and solvent B (0.1 % formic acid/80 % acetonitrile). The mass spectrophotometer was operated in full MS scan mode under positive ionization with a resolution of 60,000, a normalized automatic gain control (AGC) target of 300 %, and a mass scan range of 200–2000 *m/z*. The top 10 most intensive spectra were subjected to MS/MS at an Orbitrap resolution of 30,000, a 30 % collision energy of higher-energy collisional dissociation (HCD), a normalized AGC target of 100 %, an isolation window of 2.0 m/z, and a maximum injection time of 200 ms.

The data was analysed using Proteome Discoverer (version 2.5) with the Sequest HT searching algorithm against the pea proteome database at Uniprot (https://www.uniprot.org/taxonomy/3888). The parameters for database searching were set as follows: trypsin as the used enzyme, peptide length of 5–50, maximum missed cleavage of 2, precursor mass tolerance of 10 ppm, and fragment mass tolerance of 0.05 Da. The oxidation of methionine (+15.996 Da) was set as the dynamic modification, and the static modification was cysteine carbamidomethylation (+57.021 Da). To simplify the output, the protein database was reduced based on the results of an initial search using the complete database containing 64,000 sequences. To reduce the complexity of protein identification resulting from the poor annotation of the pea protein database, the top 350 protein hits were aligned, and sequences with >98 % similarity were clustered, retaining 145 (in the Appendix) unique sequences with the highest annotation levels. To assess the confidence of peptide-spectrum matches (PSMs), Percolator was applied to calculate the q-value and probability factor for the identified peptide-spectrum match.

2.8. *Z-potential*

Pea flour (0.4 g) was dispersed in 0.3 M NaCl at a 1:50 (*w/v*) ratio and stirred overnight at room temperature using a roller mixer (Buch & Holm A/S, Denmark). The suspension was centrifuged (Sigma 3 K15, Sigma Laborzentrifugen Gmbh, Ostrode, Germany) at 8000 rpm for 20 min at 10 °C. The supernatant was collected, and protein concentration was estimated with a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, USA) using absorbance at 280 nm. Zeta potential was determined with Zetasizer Nano ZSP (Malvern Panalytical Ltd., Malvern, UK) at 25 °C. Sample solutions were diluted with 0.3 M NaCl to acquire a final protein concentration of 2 mg/mL. The pH was adjusted using 0.1 M HCl and 0.1 M NaOH. Zeta potentials were read at increments of 1.0 between pH 3 and 7 with a pH tolerance of 0.1. Each measurement was repeated three times. The isoelectric point (pl) was calculated using the linear interpolation method.

2.9. *Statistical analysis*

All experiments were independently repeated twice. Results are expressed as mean values ± standard deviations. Statistical differences were analysed using one-way ANOVA with a *t*-test at a significant level of 0.05. Data analyses were conducted using R (version 4.3.2). Pearson's correlation tests were executed with the 'stats' package. To capture compositional diversity among accessions, the partitioning around medoids (PAM) clustering algorithm was implemented using the 'cluster' package in R, based on key macronutrient contents including protein and starch.

3. Results and discussion

3.1. *Morphological characterisation through image analysis*

The morphology of pea seeds, such as shape, size, and colour, is influenced by genetic and environmental factors, and serves as an indicator of composition and seed functionality during cooking (Dueholm et al., 2024; Santos et al., 2019). Recent genomic studies have demonstrated the value of integrating phenotypic and genotypic data to map key agronomic traits. For instance, Yang et al. (2022) used resequencing data from 118 pea accessions to re-identify Mendel's genetic *loci* controlling seed shape, showing strong concordance between phenotypic variation and population structure based on genome-wide SNP and structural variation analyses. Therefore, understanding the morphological traits of pea accessions can be used to map their compositions and functional properties. As shown in Fig. 2, significant morphological variation was observed among the 1942 accessions, which is attributed to differences in genetic specificity and the environmental growing conditions (S. K. Chen et al., 2023). Notably, accessions labelled as

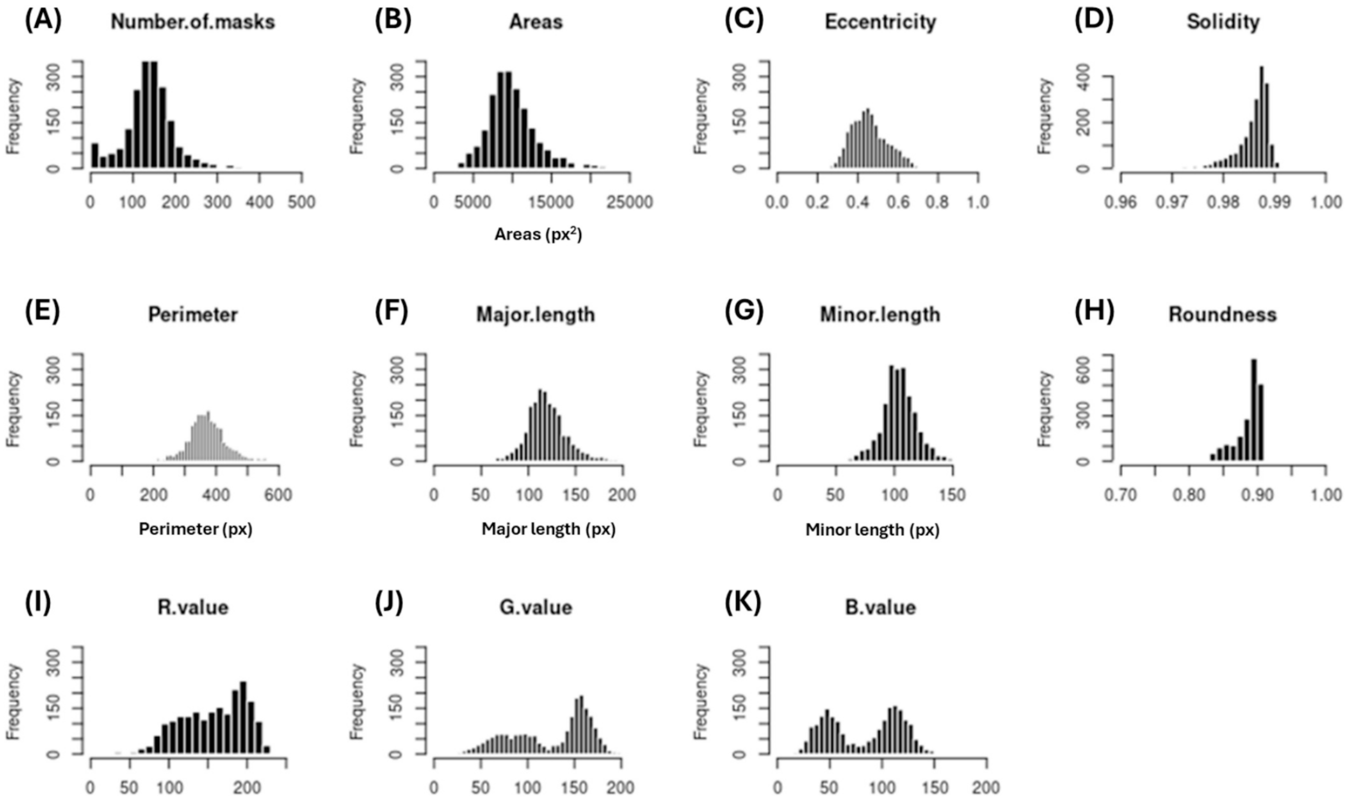

Fig. 2. Distribution histogram of the number of masks (A), geometric features (B—H), and colour information (I-J) extracted from 1942 pea accessions. Only mean values are shown.

"landrace" showed greater phenotypic variability in area, perimeter, and RG values compared to "cultivar" accessions, as indicated by their higher standard deviation (data not shown). This phenotypic diversity suggests a broader genetic base in landrace accessions, which may contribute to variation in nutritional compositions.

To efficiently quantify morphological diversity among the pea accessions, the Segment Anything Model (SAM) was used to automatically generate segmentation masks from standardized images, enabling precise extraction of size, shape, and colour features. Each mask was defined as a single-coloured enclosed shape covering one seed (Fig. 1. B). In Fig. 2. A, the number of masks indicates the number of pea seeds in an image identified by the algorithm. Generally, fewer masks per image indicate larger seed sizes. Most pea accessions produced medium-sized kernels, with a few exhibited very small or large ones. The size of pea seeds (area, perimeter, major, and minor length) followed a normal distribution. Eccentricity values were concentrated around 0.5, indicating a pronounced ratio of the focal distance to the major length. This suggests that the seeds are slightly elongated ellipses rather than perfectly spherical. The solidity was computed to quantify the convex area of the cross-section, under the assumption that values closer to 1 indicate fewer dents and a less wrinkled surface, and vice versa. This is relevant since previous studies have demonstrated significant differences in the nutritional composition, especially the amylose-to-amylopectin ratio, between smooth and wrinkled peas (Sun et al., 2023). High roundness means that the relationship between the area and perimeter of the mask is close to that of a circle.

The colour of pea coats can be homogeneous, variegated, or speckled. The RGB values from different pea accessions showed distinct distribution patterns, reflecting the pigmentation variations in pea coats (Fig. 2, I—K). The R-value displayed a unimodal distribution with relatively uniform frequencies, indicating that most pea seed coats contain a certain degree of red hue. In contrast, the G and B values exhibited bimodal distributions, suggesting two distinct groups in terms of green and blue intensity. At lower R values, both G and B values were relatively low, corresponding to darker hues such as brownish or reddish tones. These darker colours likely indicate a high accumulation of phenolic compounds (Quilichini et al., 2022). Conversely, higher R, G, and B values were associated with lighter seed coats, including yellow and greenish hues, typically low in proanthocyanidins. This analysis highlights the presence of not only commonly studied yellow and green peas, but also a significant proportion of dark-coloured accessions, such as brown and black peas, which have been relatively understudied (Sharma & Gupta, 2023). The second peak observed in the bimodal distributions of G and B values corresponds to these darker pea accessions (Fig. 2, J-K). Similar RGB distribution patterns and relationships between RGB values and seed coat colour intensity have been reported in other legumes, where lower RGB values corresponded to darker colours and higher values to lighter ones (Lay et al., 2024), supporting the robustness of this digital phenotyping approach.

It is hypothesised that the morphological differences observed among pea accessions are linked to their starch and protein content and composition. Therefore, these morphological traits provide a valuable basis for the strategic selection of genotypes aimed at enhancing specific functional and nutritional qualities. To effectively utilise these morphological traits in breeding programs, accurate and high-throughput phenotyping methods are required. Recent advances in image processing and machine learning offer powerful tools for quantifying these traits with high precision.

3.2. Establishment of selection criteria

The dataset provided by NordGen includes complete records of thousand kernel weight (TKW) and protein content for 1448 accessions out of the 1942 accessions, as only these accessions have both traits fully documented. To identify the most relevant selection criteria, correlation and principal component analysis (PCA) were conducted using protein content, TKW, and image features. The correlation coefficients are shown in Fig. 3. A. Strong correlations were observed among size-related variables (perimeter, major and minor length) and colour parameters (RGB values), potentially leading to multicollinearity issues. Therefore, it is necessary to reduce the number of variables used as selection criteria to ensure more robust and reliable results.

It has been found that TKW showed a Pearson correlation of 0.68–0.70 with seed size (perimeter, area, major and minor length), which is expected since larger seeds generally accumulate more dry matter. This accumulation contributes to higher yield, a critical parameter in modern agriculture. In contrast, TKW and protein content exhibit a modest negative correlation (coefficient of −0.15), suggesting a trade-off between carbohydrate accumulation and protein content. This phenomenon is well-documented in legumes, where increased carbohydrate deposition often occurs at the expense of protein concentration due to competition for carbon and nitrogen resources during seed development (Golombek et al., 2001; Morin et al., 2022). Therefore, TKW could serve as a valuable criterion for selecting pea accessions, particularly when targeting specific compositional traits such as carbohydrate content. PCA was further performed to evaluate the relationships among all variables. In Fig. 3. B, each arrow represents a variable, its direction and length show the relationship between variables and their principal components. The contribution of a variable is represented by a colour gradient, with red indicating the largest contribution and blue indicating the lowest contribution. Dim1 (42.3 %) predominantly captures size-related variation, especially perimeter, areas, major and minor axis length, while Dim2 (25.6 %) captures seed shape and colour parameters such as roundness, eccentricity, and RGB values. The angles between the arrows represent correlation strength. For instance, the acute angle between TKW and perimeter confirmed their high positive coefficient, consistent with the results in Fig. 3. A. Protein content is located near the coordinate origin of the principal components 1 and 2, indicating a weak correlation with morphology. Nevertheless, it remains a crucial selection criterion due to its nutritional significance. Interestingly, RGB values show minimal correlation with seed size and TKW, as reflected by nearly orthogonal vectors. While a previous study on 16 pea cultivars found that cultivars with yellow seeds had higher weight and larger diameter compared to those with green seeds (Guindon et al., 2021). This trend was not observed across the larger and more diverse set of 1448 accessions in this study. Despite their weak correlation with geometric traits or protein content, pea coat colour remains important due to its association with nutritional and functional components like phenols and flavonoids (Zhong et al., 2018). Based on this analysis, four variables, i.e., protein content, TKW, perimeter, and G-value were selected as key selection criteria for downstream analysis of the NordGencollection. These traits capture essential aspects of compositional and functional diversity relevant to pea protein utilisation.

3.3. Summary of representative accessions

The pea seeds from 51 accessions selected from NordGen are listed in Table 1. They varied in quantity, with total seeds weight per accession ranging from 0.9 g to 10.9 g. The diversity of traits among the selected samples was visually validated (Fig. 4). While the NordGen dataset comprises 1942 pea accessions, complete records for both thousand kernel weight (TKW) and protein content are only available for 1448 accessions due to missing values. The PCA plot illustrates the distribution of the selected samples (black dots) relative to the entire dataset (green dots). PC1 accounts for 45.4 % of the total variance and is primarily driven by KTW and perimeter, while PC2 accounts for 28.1 % of the variance and is mainly influenced by protein content. The selected samples are evenly distributed across the principal component space, confirming a comprehensive representation of the diversity spectrum present in the full dataset. The album of the 51 selected pea accessions further highlights the range of phenotypic diversity, particularly in seed coat colour, underscoring the effectiveness of the selection process in capturing visually observable variation.

The distribution histograms of the 51 selected pea accessions across the four criteria are shown in Fig. 5. Overall, the perimeter and G-value distributions show similar peak shapes to those in Fig. 2 E and Fig. 2 J, respectively. Although the shoulders appear smoother, indicating a less pronounced central tendency. Both protein content and TKW display normal distributions, effectively capturing both high and low extremes. These results demonstrate that the algorithm-assisted sampling strategy successfully provides robust representativeness for the key variables. Further investigation of outliers observed in the histograms was conducted to validate their classification. This included verification of their morphological features and taxonomic identity to confirm they belong to *Pisum sativum*. However, further genetic analysis may be required to rule out the possibility of hybridisation or mislabelling.

3.4. Chemical compositions and correlation with image features

Table 1 summarises the main compositions of the 51 accessions, the isoelectric point, and their taxonomy information. Notably, the wrinkle/

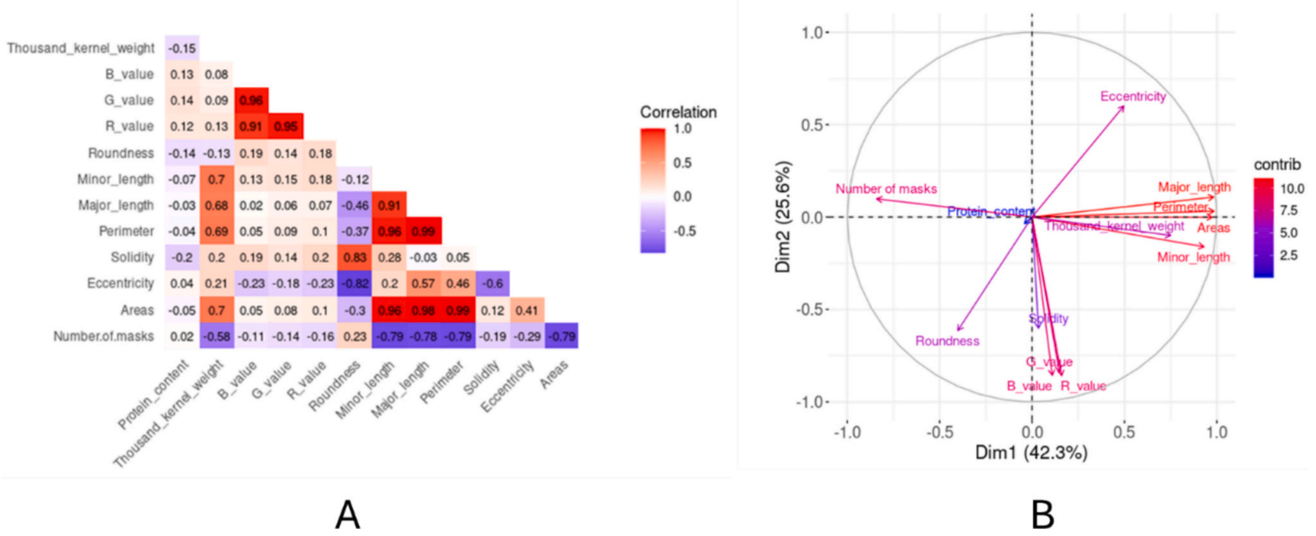

Fig. 3. (A) Correlation heatmap showing the relationships between thousand kernel weight, documented protein content, and 13 image features of 1448 pea accessions. (B) PCA analysis illustrates these variables' contributions to the first two principal components (Dim1 and Dim2).

Table 1
Overview of the 51 selected pea accessions and their compositions. Values are expressed as mean ± standard deviation based on duplicate measurements.

NordGen ID	Taxonomy	Name	Wrinkle/ Smooth	Protein content (%, on dry basis)	Starch content (%, on dry basis)	Amylose content (%, on dry starch basis)	Other components (%)	Isoelectric point
NGB103559	*Pisum abyssinicum* A. Braun	WBH 3559	S	24.9 ± 0.5	41.8 ± 0.2	27.5 ± 3.1	33.3 ± 0.7	4.3
NGB103067	*Pisum sativum* L.	WBH 3067	S	17.1 ± 0.1	43.6 ± 0.2	30.3 ± 0.8	39.3 ± 0.3	4.0
NGB103619	*Pisum sativum* L.	Braychangma	S	18.4 ± 0.1	44.2 ± 0.5	35.3 ± 1.7	37.4 ± 0.6	3.9
NGB102987	*Pisum sativum* L.	WBH 2987	S	20.6 ± 0.5	41.1 ± 0.0	36.3 ± 4.0	38.3 ± 0.5	4.6
NGB103724	*Pisum sativum* L.	WBH 3724	S	20.9 ± 0.8	36.6 ± 1.3	28.8 ± 1.4	42.5 ± 2.1	4.1
NGB103579	*Pisum sativum* L.	WBH 3579	S	21.2 ± 0.2	42.3 ± 1.3	15.5 ± 1.9	36.5 ± 1.5	4.3
NGB103852	*Pisum sativum* L.	Jo 1068	S	21.3 ± 0.2	45.3 ± 0.7	31.2 ± 1.0	33.4 ± 0.9	4.6
NGB105881	*Pisum sativum* L.	Chlorotica	S	21.4 ± 0.2	44.7 ± 0.1	29.3 ± 1.7	33.9 ± 0.3	4.0
NGB106020	*Pisum sativum* L.	Medicagoides	S	22.5 ± 0.9	42.7 ± 0.7	27.7 ± 2.8	34.8 ± 1.6	4.4
NGB106104	*Pisum sativum* L.	Calyx carpellaris	S	22.5 ± 0.2	42.7 ± 0.8	43.6 ± 0.0	34.8 ± 1.0	4.1
NGB103607	*Pisum sativum* L.	WBH 3607	W	22.6 ± 0.1	33.3 ± 0.1	36.4 ± 4.7	44.1 ± 0.2	3.9
NGB103758	*Pisum sativum* L.	WBH 3758	W	22.7 ± 0.1	29.6 ± 0.3	36.7 ± 2.1	47.7 ± 0.4	3.8
NGB101228	*Pisum sativum* L.	WBH 1228	W	22.7 ± 0.1	43.0 ± 2.5	44.3 ± 0.7	34.3 ± 2.6	4.0
NGB105898	*Pisum sativum* L.	Chlorotica	S	22.8 ± 0.1	46.0 ± 2.0	28.0 ± 0.3	31.2 ± 2.1	3.9
NGB102153	*Pisum sativum* L.	Falensky-42	S	22.9 ± 0.0	43.2 ± 0.5	30.6 ± 4.7	33.9 ± 0.5	4.4
NGB101776	*Pisum sativum* L.	WBH 1776	S	23.0 ± 0.8	38.8 ± 0.3	29.0 ± 0.8	38.2 ± 1.1	4.2
NGB103624	*Pisum sativum* L.	WBH 3624	W	23.1 ± 0.2	21.0 ± 1.1	48.1 ± 2.6	55.9 ± 1.3	4.3
NGB103729	*Pisum sativum* L.	WBH 3729	S	23.2 ± 0.4	36.6 ± 0.7	27.2 ± 4.1	40.2 ± 1.1	4.2
NGB102115	*Pisum sativum* L.	Cobri	S	23.2 ± 0.5	42.0 ± 0.9	27.8 ± 2.5	34.8 ± 1.4	3.9
NGB106021	*Pisum sativum* L.	Medicagoides	S	23.4 ± 0.0	41.6 ± 0.1	25.6 ± 1.0	35.0 ± 0.1	4.6
NGB103036	*Pisum sativum* L.	WBH 3036	S	23.4 ± 0.2	41.4 ± 0.9	28.5 ± 1.6	35.2 ± 1.1	4.2
NGB101519	*Pisum sativum* L.	WBH 1519	S	23.7 ± 1.0	38.9 ± 1.4	35.7 ± 0.3	37.4 ± 2.4	3.8
NGB106138	*Pisum sativum* L.	WBH 6138	S	23.8 ± 0.2	43.1 ± 0.3	29.2 ± 0.3	33.1 ± 0.5	4.2
NGB105895	*Pisum sativum* L.	Subtus-incerata	S	24.2 ± 0.2	40.5 ± 0.9	25.5 ± 4.4	35.3 ± 1.1	3.9
NGB103751	*Pisum sativum* L.	WBH 3751	S	24.4 ± 0.2	35.1 ± 1.2	32.5 ± 2.7	40.5 ± 1.4	3.9
NGB103621	*Pisum sativum* L.	WBH 3621	S	24.7 ± 0.6	41.9 ± 0.2	23.9 ± 0.4	33.4 ± 0.8	4.0
NGB101515	*Pisum sativum* L.	WBH 1515	W	24.8 ± 0.4	28.3 ± 0.7	64.3 ± 0.7	46.9 ± 1.1	3.8
NGB105931	*Pisum sativum* L.	Supra-incerata	S	24.8 ± 0.5	42.9 ± 0.3	21.3 ± 1.3	32.3 ± 0.8	4.0
NGB101535	*Pisum sativum* L.	WBH 1535	S	24.9 ± 0.1	37.5 ± 0.4	35.0 ± 1.2	37.6 ± 0.5	4.2
NGB101496	*Pisum sativum* L.	WBH 1496	S	25.4 ± 0.1	39.9 ± 1.1	22.7 ± 0.0	34.7 ± 1.2	4.0
NGB105790	*Pisum sativum* L.	chlorotica	S	25.5 ± 0.2	38.5 ± 1.9	34.4 ± 3.3	36.0 ± 2.1	4.3
NGB103095	*Pisum sativum* L.	WBH 3095	W	25.7 ± 0.0	30.0 ± 0.7	59.1 ± 2.3	44.3 ± 0.7	4.0
NGB105039	*Pisum sativum* L.	longo-internodium	S	25.9 ± 0.1	41.3 ± 0.7	27.6 ± 0.1	32.8 ± 0.8	4.1
NGB106080	*Pisum sativum* L.	WBH 6080	S	26.0 ± 0.4	40.0 ± 1.2	29.6 ± 1.1	34.0 ± 1.6	4.2
NGB105345	*Pisum sativum* L.	Variomicromaculata	S	26.1 ± 0.5	38.8 ± 0.4	24.7 ± 2.0	35.1 ± 0.9	3.9
NGB103800	*Pisum sativum* L.	WBH 3800	W	26.3 ± 0.1	33.7 ± 0.9	29.8 ± 0.8	40.0 ± 1.0	4.4
NGB100756	*Pisum sativum* L.	WBH 756	W	26.3 ± 1.2	28.4 ± 0.7	73.3 ± 0.7	45.3 ± 1.9	4.2
NGB106110	*Pisum sativum* L.	Desynaptic	S	26.4 ± 0.1	39.3 ± 0.1	27.1 ± 0.6	34.3 ± 0.2	4.2
NGB106006	*Pisum sativum* L.	Desynaptic; Homozygous	S	26.4 ± 0.1	39.5 ± 0.7	26.7 ± 0.6	34.1 ± 0.8	3.9
NGB103761	*Pisum sativum* L.	WBH 3761	W	27.3 ± 0.5	28.2 ± 0.8	28.5 ± 1.1	44.5 ± 1.3	3.7
NGB105048	*Pisum sativum* L.	Chlorotica	S	27.8 ± 0.5	39.5 ± 0.8	33.3 ± 1.1	32.7 ± 1.3	4.6
NGB105447	*Pisum sativum* L.	Xantha	W	29.2 ± 0.6	33.2 ± 0.5	23.7 ± 1.9	37.6 ± 1.1	3.7
NGB100464	*Pisum sativum* L. subsp. *sativum*	Apollo II	W	17.8 ± 0.3	31.7 ± 0.6	62.4 ± 9.9	50.5 ± 0.9	4.2
NGB101463	*Pisum sativum* L. subsp. *sativum*	Sigyn	W	22.9 ± 0.2	28.9 ± 0.4	69.4 ± 0.4	48.2 ± 0.6	4.4
NGB103867	*Pisum sativum* L. subsp. *sativum*	N/A	S	22.5 ± 0.2	44.9 ± 0.4	29.3 ± 1.4	32.6 ± 0.6	4.0
NGB100657	*Pisum sativum* L. subsp. sativum	Balder	W	24.6 ± 0.8	28.1 ± 0.1	64.3 ± 0.8	47.3 ± 0.9	4.4
NGB102764	*Pisum sativum* L. subsp. *sativum*	Olympia	W	24.9 ± 0.2	30.2 ± 0.1	60.2 ± 4.5	44.9 ± 0.3	3.7
NGB102663	*Pisum sativum* L. subsp. *sativum*	WBH 2663	W	29.6 ± 0.5	22.5 ± 1.1	72.3 ± 6.0	47.9 ± 1.6	4.5
NGB102052	*Pisum sativum* L. subsp. sativum var. *arvense* (L.) Poir.	WBH 2052	S	21.7 ± 1.0	48.1 ± 2.1	18.0 ± 0.8	30.2 ± 3.1	4.6
NGB103371	*Pisum sativum* L. *subsp. sativum* var. *arvense* (L.) Poir.	Fregero	S	22.0 ± 0.6	40.4 ± 0.1	32.4 ± 2.5	37.6 ± 0.7	4.6
NGB101596	*Pisum sativum* L. subsp. *sativum* var. *arvense* (L.) Poir.	WBH 1596	S	24.4 ± 0.1	39.4 ± 0.6	21.8 ± 0.6	36.2 ± 0.7	4.3

N/A: Not available.

smooth phenotype classification presented in the table was determined based on visual inspection. The protein content among accessions ranged from 17.1 to 29.6 % and the starch content from 21.0 to 48.1 %, covering a broader range than the commonly reported 20–25 % for peas (Shen et al., 2022). Compared to earlier reports, the present study analysed a more extensive and diverse set of accessions, allowing a broader assessment of compositional variation.

Amylose content varied considerably among accessions (15.5–69.4 %), influenced by both genetic and environmental factors. Variations in starch-branching enzyme I (SBEI) activity contribute to differences in

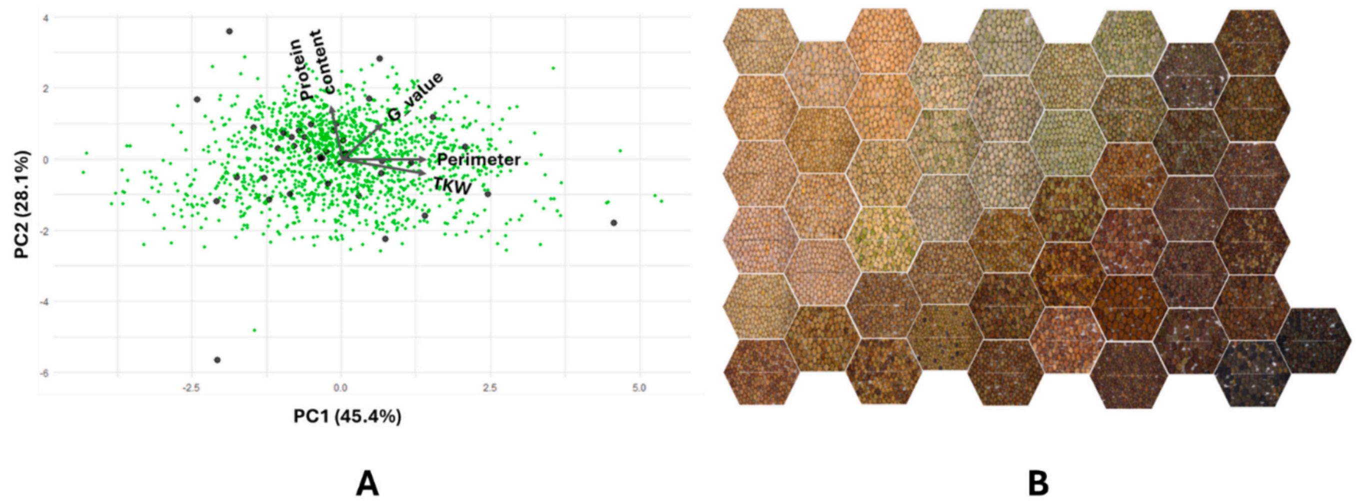

Fig. 4. A: Principal Component Analysis (PCA) plot showing the distribution of the 51 selected pea accessions (black dots) within the total dataset of 1942 accessions (green dots). The arrows indicate the contributions of key variables (Protein content, Perimeter, TKW) to the principal components. B: Image album showcasing the phenotypic diversity among the 51 selected pea accessions. Photos provided by NordGen. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

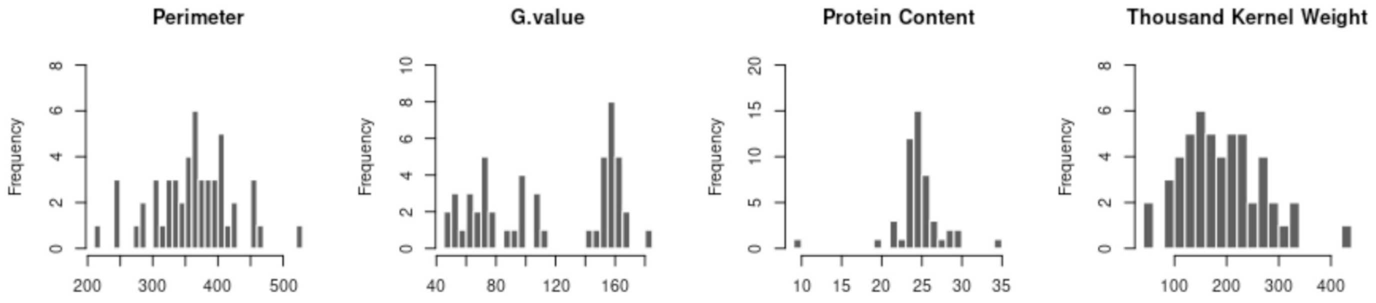

Fig. 5. Distribution histograms of the selected 51 pea accessions on the four selection criteria.

amylose and amylopectin ratios, while other starch biosynthesis enzymes and environmental conditions further modulate starch composition. Beyond starch synthesis, differences in sucrose accumulation and osmotic pressure affect seed structure, with some accessions retaining more water, leading to greater shrinkage upon drying and a wrinkled appearance (Cheng et al., 2024; Moreau et al., 2022). Notably, not all wrinkled accessions had higher amylose content compared to smooth accessions. For instance, the wrinkled accession NGB103800 had lower amylose content than NGB101535, which has smooth seeds. While the wrinkled phenotype in peas is most commonly attributed to the rugosus (*rr*) mutation, other factors can also contribute. Mutations in other starch-related enzymes, modifier genes, or complex epistatic interactions can influence starch biosynthesis and lead to a wrinkled appearance (Daba et al., 2024). Additionally, environmental factors such as water stress or temperature fluctuations during seed development can affect moisture retention and starch structure, contributing to variability in wrinkling (Rayner et al., 2017). Therefore, the lower amylose content observed in NGB103800, despite its wrinkled phenotype, could be influenced by these additional genetic and environmental factors, particularly given its landrace origin, which is associated with a more diverse genetic background.

The correlation analysis between image features and chemical compositions across the 51 pea accessions is presented in Fig. 6. TKW showed negligible correlations with protein content, total starch, amylose content, and other components, suggesting that TKW primarily reflects seed size and dry matter rather than compositional traits. Therefore, compositional differences among accessions are not determined by TKW. Previous research has shown that the correlation between protein and starch content is not consistent, with studies reporting both negative correlations and no significant relationships (Daba & Morris, 2022). Analysis of the 51 pea accessions in this study revealed a strong negative correlation between starch content and amylose content (coefficient: −0.72) as well as other components (coefficient: −0.87), but no significant correlation with protein content. This suggests the possibility of selecting pea accessions with high levels of both protein and starch, as exemplified by accessions such as NGB105048 and NGB106006. Additionally, amylose content showed a positive correlation with other components (coefficient: 0.73) but no correlation with protein content. Dueholm et al. (2024) analysed 19

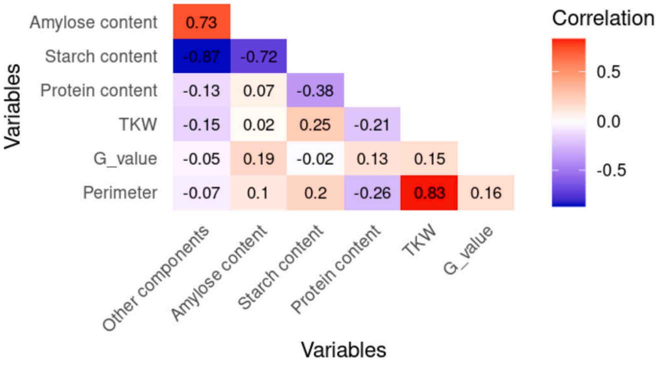

Fig. 6. Heatmap of correlation between image features and chemical compositions across 51 pea accessions.

smooth and 5 wrinkled pea accessions and found no correlation between amylose and protein content. However, when the 5 wrinkled accessions were excluded, a strong negative correlation (−0.7) was observed. This shift indicates that phenotypic classification (smooth vs. wrinkled) can markedly influence observed compositional relationships. Considering the broad variation among the 51 genetically diverse pea accessions analysed in this study, the results suggest that correlations between starch, protein, and other traits are context-dependent and may be influenced by multiple factors, including genetic background, seed phenotype, and environmental conditions.

Interestingly, solidity exhibits a strong negative correlation with amylose content (coefficient: −0.7) and positively correlated with starch content (coefficient: 0.73). Roundness and eccentricity also showed correlations with amylose and starch content, though their coefficients were lower than those of solidity. In contrast, neither seed size (areas, perimeter, major length, and minor length) nor colour (RGB values) demonstrated strong correlations with the studied compounds. In summary, solidity is a geometric feature capable of characterising seed wrinkling, as it showed strong correlations with total starch, amylose content, and other components. However, the predictive power of other image features for protein content was limited. These findings suggest that geometric features extracted from images are more indicative of carbohydrate-related traits in peas.

3.5. Comparative starch analysis via WAXS

Starch played a critical role in determining the textural and functional properties of food ingredients. WAXS is a technology capable of revealing structural information with minimal sample volume. It was used to investigate the crystalline structure of starch in all pea accessions. The X-ray scattering patterns revealed prominent peaks at 2θ 15.0°, 18°, 20.2°, and 22.7°, characteristic of the A-type crystalline polymorph commonly found in starches. Smooth pea accessions exhibited sharper and more defined peaks, indicating higher starch crystallinity, while wrinkled pea accessions showed broader and less intense peaks, suggesting reduced crystallinity and a more amorphous structure (AL-Ansi et al., 2021). These differences were further quantified, with smooth peas showing an average crystallinity of 12.9 % and wrinkled peas 12.1 %, a statistically significant difference ($P < 0.05$; Fig. 7). Both values were lower than the previously reported ranges of 23.8–31.3 % for smooth peas and 19.2–20.8 % for wrinkled peas (Cheng et al., 2024; Shi et al., 2023), likely due to differences in genetic background, limited sample size, and variations in the calculation software and methods applied.

The lower crystallinity observed in wrinkled peas was consistent with their higher amylose content, as amylose inhibited the formation of tightly packed crystalline amylopectin double helices (Shi et al., 2023). These structural differences were likely to influence functional properties such as water/oil absorption, gelatinization, and nutritional profiles, attributes that were critical for the development of plant-based food products (Ren et al., 2021). Grouping information of features with large differences is displayed in Fig. 7, further illustrating the structural distinctions between smooth and wrinkled pea accessions.

3.6. Comparative protein analysis via LC-MS/MS

The protein subunit profiles of 51 pea accessions were characterised by LC-MS/MS analysis. A total of 1161 proteins (see Appendix) were detected across all samples. To reduce the complexity of the protein identification because of poor annotation of the pea protein database, the top 350 protein hits were aligned, and sequences with >98 % similarity were clustered, retaining 145 unique sequences with the highest annotation levels (see Appendix). PCA was performed to visualise the variation in these 145 protein profiles across the 51 pea accessions. PC1 accounted for 52.73 % of the variance, while PC2 explained 18.12 %. The loading plot revealed that legumin (A2, J, B), vicilin, and convicilin contributed significantly to the variation among accessions (Fig. 8. A). The variation in the PC1 direction is primarily driven by legumin A2 and the legumin J isoform (A0A9D4W585), whereas other isoforms of legumin J (P05692) and (A0A9D5A4B1) are located much closer to the centre. Illustrating a larger variation of legumin J isoform (A0A9D4W585) across the 51 pea accessions compared with the two

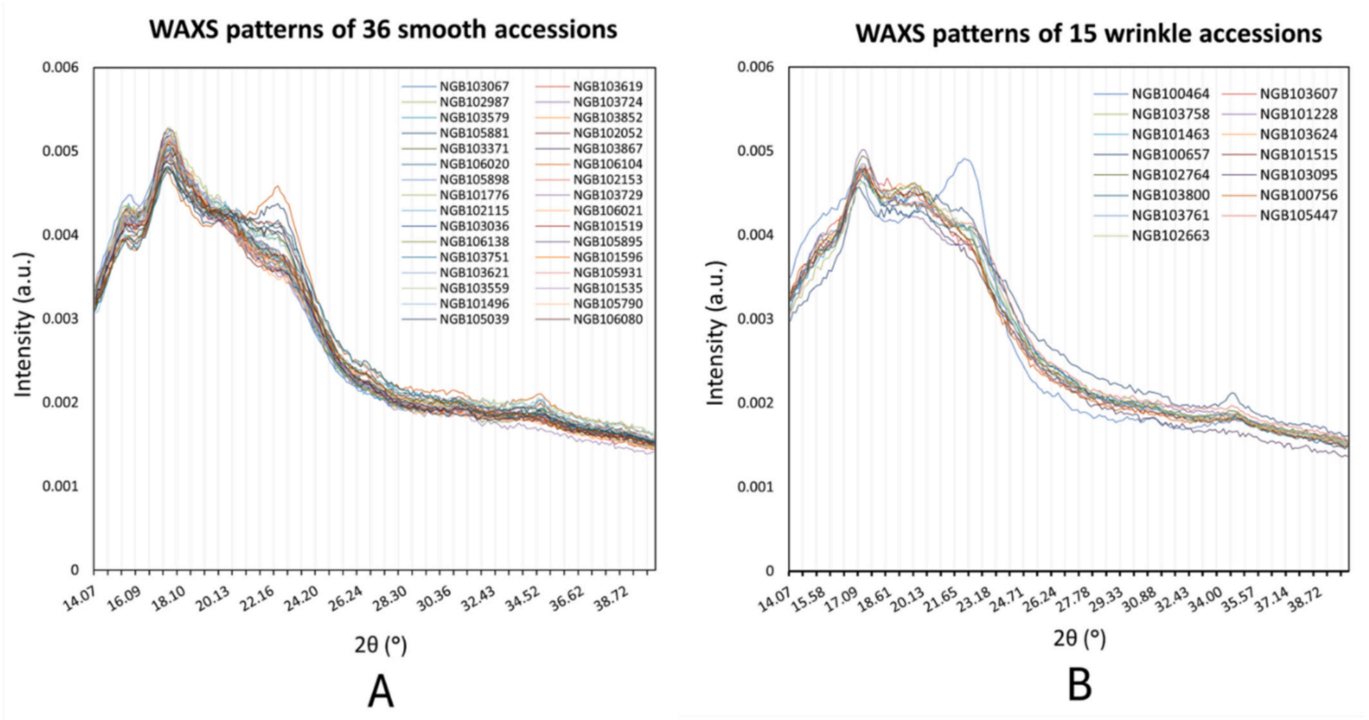

Fig. 7. WAXS patterns show the grouping information, differentiating 36 smooth pea accessions (left) and 15 wrinkled pea accessions (right). The overall scattering pattern and the peak at 2θ 15.0° are good indicators of distinguishing smooth from wrinkled peas.

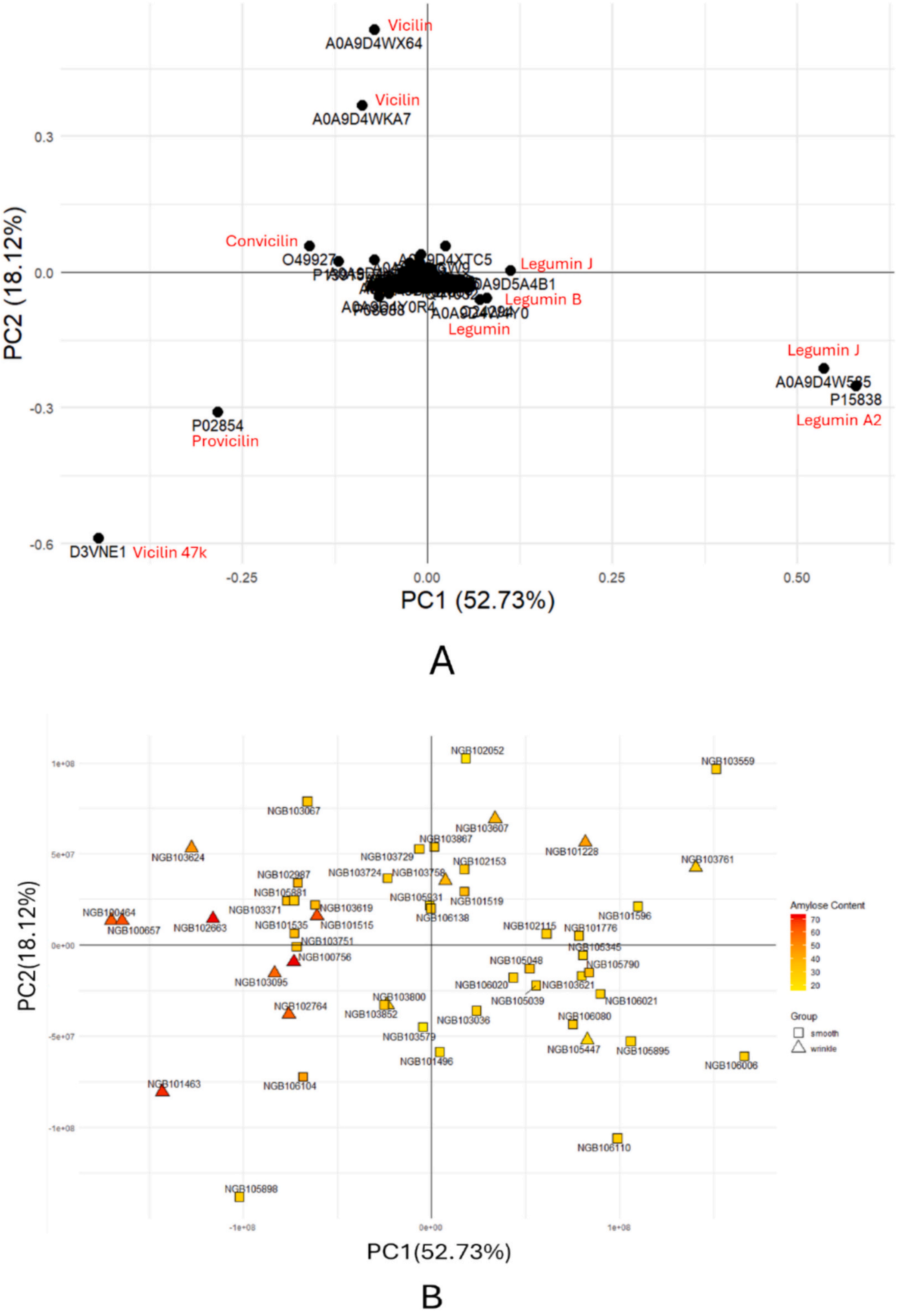

Fig. 8. A: PCA loading plot of protein subunits across 51 pea accessions. B: PCA score plot coloured by amylose content. Wrinkled and smooth accessions are marked with triangles and squares, respectively.

other legumin J isoforms (P05692) and (A0A9D5A4B1). The variation in the PC2 direction is mainly dominated by the variation of different types of vicilin indicating that there is a natural variation in vicilin types like vicilin, con-vicilin, pro-vicilin, and vicilin 47 k that is to some extent independent of the variation of legumin A2 and legumin J. The PCA score plot (Fig. 8. B) illustrated a relatively equal distribution of the 51 accessions along the PC1 and PC2 directions, with no significant grouping. Colouring of the score plot according to amylose content indicated a relationship between the variation of amylose content and the variation in the PC1 direction, and it was found that vicilin content was related to amylose content. By combining the loading plot and score plot, it was inferred that legumins (A2 and J) prevailed in smooth peas, while the different types of vicilins were more prevalent in wrinkled ones, which was consistent with the result from Daba et al. (2024). This pattern can be explained by the reduced vicilin content in wrinkled peas, resulting from increased osmotic stress in the high-sugar environment during seed development, which in turn led to the instability of legumin mRNA (Daba & Morris, 2022). Compared to the limited variation observed across eight cultivars in Vreeke et al. (2023), the 51 pea accessions analysed in this study exhibited a broader dynamic range in globulin subunit composition. Notable differences were found in the abundance of vicilin and legumin, and these variations were associated with phenotypic traits such as starch composition. This suggested a greater protein diversity, likely resulting from the wider genetic background covered in the current panel. Supporting this, Kreplak et al. (2019) annotated 12 legumin and 9 vicilin genes in the pea reference genome, revealing a substantial expansion of these storage protein gene families. This genomic evidence supports the observed variation in legumin and vicilin abundance across accessions, suggesting that differences in protein subunit profiles may partly reflect divergence in gene copy number or expression. Meanwhile, Rayner et al. (2024) found that the removal of vicilin by deletion of the corresponding genes did not reduce the protein concentration in mature pea seeds. This suggested that protein composition was linked to starch content, potentially through shared developmental pathways, while total protein content remained independent of starch composition. This observation supported the hypothesis that protein accumulation mechanisms are regulated separately from carbohydrate metabolism. In addition to globulin subunits, 12 proteins were annotated as protease inhibitors. While these were detected in many accessions, their abundance was relatively low, and their contribution to PCA variation was minimal. Therefore, they were not a primary focus in this study, which emphasises storage protein composition and functionality.

Legumin and vicilin, also referred to as 11S and 7S globulin fractions, respectively, account for approximately 32–77 % of the total protein in peas (Gravel et al., 2024). These globulins play a crucial role in determining protein functionality, including solubility, gelation, and nutritional properties. Previous studies have found that a higher 7S/11S ratio generally promotes better gelling properties, however, certain vicilin subunits may inhibit gelation (Husband et al., 2024). By mapping pea protein compositions alongside other key components such as starch, it becomes feasible to manipulate the legumin-to-vicilin and amylose-to-amylopectin ratios through careful accession selection. Consequently, protein concentrates with desired functional properties can be obtained with minimal reliance on intensive fractionation processes.

3.7. ζ-potential

To explore the colloidal stability of pea protein extracts, ζ-potential was measured across five pH values for all 51 accessions. The distinct patterns observed indicated substantial differences in buffering capacities (see supplementary data). To assess whether these trends corresponded to protein composition, the average isoelectric point (pI) of each sample was calculated based on LC-MS/MS subunit profiles (Table 1) (Helmick et al., 2021). Moreover, the correlation between ζ-potential at different pH and legumin-to-vicilin ratio was weak (e.g., a coefficient of −0.24 at pH 4). This indicates that surface charge behaviour is not solely determined by protein sequence composition, non-protein components such as phytic acid and polysaccharide complexes may also affect protein solubility (Tanger et al., 2020).

At pH 7, the ζ-potential of proteins from different pea accessions ranged from −11 mV to −3 mV. The range even exceeded the differences caused by various extraction methods applied to a single pea cultivar (Tanger et al., 2020), highlighting the importance of leveraging natural variation among accessions. However, it remains notably lower than the commonly recognised ±30 mV threshold, beyond which colloidal dispersions are generally considered stable. The relatively low ζ-potential values observed in this study suggested that pea protein dispersions may exist in a semi-stable state, where hydrophobic interactions, hydrogen bonding, and van der Waals forces could play a more significant role in aggregation. The isoelectric points ranged from 3.7 to 4.6 among the 51 accessions (see Table 1), providing a useful reference for identifying pea accessions with ζ-potential trends that remain stable within target pH ranges. Such accessions may be more suitable for food processing applications requiring stable protein dispersions.

4. Conclusions

This study demonstrates the potential of integrating image-based phenotyping and minimally refined compositional analysis to characterise and utilise pea germplasm for food applications. By extracting geometric and colour features from 1942 pea accessions and selecting a representative subset using clustering based on protein, starch, and morphology, extensive diversity in both phenotypic and compositional traits was revealed. Algorithm-assisted selection expanded the coverage of protein and starch content compared to previous reports. Notably, solidity emerged as a key morphological marker strongly correlated with starch composition, while WAXS confirmed structural differences in starch crystallinity that underscore the functional relevance of these traits. Moreover, protein and starch contents exhibited negligible correlation.

Proteomic analysis identified legumin and vicilin as the primary contributors to variation among the 51 pea accessions, with legumin predominating in smooth and vicilin in wrinkled accessions. Variations in protein subunit composition and other compounds, such as phytic acids, also significantly influenced zeta potential and isoelectric points under different pH conditions, which formed the basis for selecting accessions according to their genetic variants.

Future research should focus on refining selection criteria for representative accessions based on key compositional traits while addressing the constraints of limited seed availability. Advancements in non-destructive phenotyping and high-throughput compositional analysis will be critical for overcoming these constraints and enabling efficient utilisation of genebank diversity. By harnessing natural genetic variation, this approach offers a sustainable alternative to conventional protein fractionation, reducing the reliance on resource-intensive processing methods while expanding the potential of pea-based food applications.

CRediT authorship contribution statement

Qinhui Xing: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Zhi Ye:** Writing – original draft, Software, Methodology. **Bo Yuan:** Validation. **Xiaoxiao Liu:** Validation, Methodology. **Morten Arendt Rasmussen:** Writing – review & editing, Software, Methodology, Formal analysis. **Jacob Judas Kain Kirkensgaard:** Writing – review & editing, Resources, Methodology. **Michael Lyngkjær:** Writing – review & editing, Resources. **Ulrika Carlson-Nilsson:** Writing – review & editing, Resources. **Cecilia Hammenhag:** Writing – review & editing, Resources. **Rene Lametsch:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

The project is supported by the Novo Nordisk Foundation (Grant number: NNF220C0079385). Data were generated using research infrastructure at the University of Copenhagen, partially funded by FOODHAY (Food and Health Open Innovation Laboratory, Danish Roadmap for Research Infrastructure).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2025.145478.

Data availability

Data will be made available on request.

References

AL-Ansi, W., Sajid, B. M., Mahdi, A. A., Al-Maqtari, Q. A., AL-Adeeb, A., Ahmed, A., … Wang, L. (2021). Molecular structure, morphological, and physicochemical properties of highlands barley starch as affected by natural fermentation. *Food Chemistry, 356*(15), Article 129665. https://doi.org/10.1016/j.foodchem.2021.129665

Arteaga, V. G., Kraus, S., Schott, M., Muranyi, I., Schweiggert-Weisz, U., & Eisner, P. (2021). Screening of twelve pea (*Pisum sativum* L.) cultivars and their isolates focusing on the protein characterization, functionality, and sensory profiles. *Foods, 10*(4), 758. https://doi.org/10.3390/foods10040758

Brückner, S. (2000). Estimation of the background in powder diffraction patterns through a robust smoothing procedure. *Journal of Applied Crystallography, 33*(3 II), 977–979. https://doi.org/10.1107/S0021889800003617

Chen, B., Shi, Y., Sun, Y., Lu, L., Wang, L., Liu, Z., & Cheng, S. (2024). Innovations in functional genomics and molecular breeding of pea: Exploring advances and opportunities. *aBIOTECH, 5*(1), 71–93. https://doi.org/10.1007/s42994-023-00129-1

Chen, S. K., Lin, H. F., Wang, X., Yuan, Y., Yin, J. Y., & Song, X. X. (2023). Comprehensive analysis in the nutritional composition, phenolic species and in vitro antioxidant activities of different pea cultivars. *Food Chemistry: X, 17*, Article 100599. https://doi.org/10.1016/j.fochx.2023.100599

Cheng, F., Ren, Y., Warkentin, T. D., & Ai, Y. (2024). Heat-moisture treatment to modify structure and functionality and reduce digestibility of wrinkled and round pea starches. *Carbohydrate Polymers, 324*, Article 121506. https://doi.org/10.1016/j.carbpol.2023.121506

Daba, S. D., & Morris, C. F. (2022). Pea proteins: Variation, composition, genetics, and functional properties. *Cereal Chemistry, 99*(1), 8–20. https://doi.org/10.1002/cche.10439

Daba, S. D., Panda, P., Aryal, U. K., Kiszonas, A. M., Finnie, S. M., & McGee, R. J. (2024). Proteomics analysis of round and wrinkled pea (*Pisum sativum* L.) seeds during different development periods. *Proteomics, 25*, Article 2300363. https://doi.org/10.1002/pmic.202300363

Dueholm, B., Fonskov, J., Grimberg, Å., Carlsson, S., Hefni, M., Henriksson, T., & Hammenhag, C. (2024). Cookability of 24 pea accessions—Determining factors and potential predictors of cooking quality. *Journal of the Science of Food and Agriculture, 104*(6), 3685–3696. https://doi.org/10.1002/jsfa.13253

Golombek, S., Rolletschek, H., Wobus, U., & Weber, H. (2001). Control of storage protein accumulation during legume seed development. *Journal of Plant Physiology, 158*(4), 457–464. https://doi.org/10.1078/0176-1617-00357

Gravel, A., Dubois-Laurin, F., Turgeon, S. L., & Doyen, A. (2024). The role of the 7S/11S globulin ratio in the gelling properties of mixed β-lactoglobulin/pea proteins systems. *Food Hydrocolloids, 156*, Article 110273. https://doi.org/10.1016/j.foodhyd.2024.110273

Guindon, M. F., Aguero, M. G., Gatti, I., & Cointry, E. (2021). Comparative analysis of the physicochemical composition of pea cultivars. *Ciencia Tecnologia Agropecuaria, 22*(3), Article e1761. https://doi.org/10.21930/rcta.vol22_num2_art:1761

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). *Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 16000–16009. New Orleans, LA*.

Helmick, H., Hartanto, C., Bhunia, A., Liceaga, A., & Kokini, J. L. (2021). Validation of bioinformatic modeling for the zeta potential of vicilin, legumin, and commercial pea protein isolate. *Food Biophysics, 16*(4), 474–483. https://doi.org/10.1007/s11483-021-09686-8

Husband, H., Ferreira, S., Bu, F., Feyzi, S., & Ismail, B. P. (2024). Pea protein globulins: Does their relative ratio matter? *Food Hydrocolloids, 148*(Part A), Article 109429. https://doi.org/10.1016/j.foodhyd.2023.109429

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., … Girshick, R. (2023). *Segment anything. Proceedings of the IEEE international conference on computer vision, 3992–4003. Paris, France*.

Kornet, R., Yang, J., Venema, P., van der Linden, E., & Sagis, L. M. C. (2022). Optimizing pea protein fractionation to yield protein fractions with a high foaming and emulsifying capacity. *Food Hydrocolloids, 126*, Article 107456. https://doi.org/10.1016/j.foodhyd.2021.107456

Kreplak, J., Madoui, M. A., Cápal, P., Novák, P., Labadie, K., Aubert, G., … Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nature Genetics, 51*(9), 1411–1422. https://doi.org/10.1038/s41588-019-0480-1

Lay, L., Khan, W., Jo, H., Kim, S. H., & Kim, Y. (2024). Genome-wide association study on cowpea seed coat color using RGB images. *Molecular Breeding, 44*(12). https://doi.org/10.1007/s11032-024-01516-2

Lie-Piang, A., Yang, J., Schutyser, M. A. I., Nikiforidis, C. V., & Boom, R. M. (2025). Mild fractionation for more sustainable food ingredients. *Annual Review of Food Science and Technology, 14*, 473–493. https://doi.org/10.1146/annurev-food-060721

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., & Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis, 89*, Article 102918. https://doi.org/10.1016/j.media.2023.102918

Moreau, C., Warren, F. J., Rayner, T., Perez-Moral, N., Lawson, D. M., Wang, T. L., & Domoney, C. (2022). An allelic series of starch-branching enzyme mutants in pea (*Pisum sativum* L.) reveals complex relationships with seed starch phenotypes. *Carbohydrate Polymers, 288*, Article 119386. https://doi.org/10.1016/j.carbpol.2022.119386

Morin, A., Maurousset, L., Vriet, C., Lemoine, R., Doidy, J., & Pourtau, N. (2022). Carbon fluxes and environmental interactions during legume development, with a specific focus on Pisum sativum. *Physiologia Plantarum, 174*(3), Article e13729. https://doi.org/10.1111/ppl.13729

Nguyen, G. N., & Norton, S. L. (2020). Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm. *Plants, 9*(7), 1–27. https://doi.org/10.3390/plants9070817

Pandey, A. K., Rubiales, D., Wang, Y., Fang, P., Sun, T., Liu, N., & Xu, P. (2021). Omics resources and omics-enabled approaches for achieving high productivity and improved quality in pea (*Pisum sativum* L.). *Theoretical and Applied Genetics, 134*(3), 755–776. https://doi.org/10.1007/s00122-020-03751-5

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., … Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library. 33rd conference on neural information processing systems (NeurIPS 2019, Vancouver, Canada)*.

Pelgrom, P. J. M., Vissers, A. M., Boom, R. M., & Schutyser, M. A. I. (2013). Dry fractionation for production of functional pea protein concentrates. *Food Research International, 53*(1), 232–239. https://doi.org/10.1016/j.foodres.2013.05.004

Quilichini, T. D., Gao, P., Yu, B., Bing, D., Datla, R., Fobert, P., & Xiang, D. (2022). The seed coat's impact on crop performance in pea (*Pisum sativum* L.). *Plants, 11*(15), 2056. https://doi.org/10.3390/plants11152056

Rayner, T., Moreau, C., Ambrose, M., Isaac, P. G., Ellis, N., & Domoney, C. (2017). Genetic variation controlling wrinkled seed phenotypes in Pisum: How lucky was mendel? *International Journal of Molecular Sciences, 18*(6), 1205. https://doi.org/10.3390/ijms18061205

Rayner, T., Saalbach, G., Vickers, M., Paajanen, P., Martins, C., Wouters, R. H. M., … Domoney, C. (2024). Rebalancing the seed proteome following deletion of vicilin-related genes in pea (*Pisum sativum* L.). *Journal of Experimental Botany*. , Article erae518. https://doi.org/10.1093/jxb/erae518/7929985

Ren, Y., Setia, R., Warkentin, T. D., & Ai, Y. (2021). Functionality and starch digestibility of wrinkled and round pea flours of two different particle sizes. *Food Chemistry, 336*, Article 127711. https://doi.org/10.1016/j.foodchem.2020.127711

Saget, S., Costa, M., Santos, C. S., Vasconcelos, M. W., Gibbons, J., Styles, D., & Williams, M. (2021). Substitution of beef with pea protein reduces the environmental footprint of meat balls whilst supporting health and climate stabilisation goals. *Journal of Cleaner Production, 297*(15), Article 126447. https://doi.org/10.1016/j.jclepro.2021.126447

Santos, C. S., Carbas, B., Castanho, A., Vasconcelos, M. W., Vaz Patto, M. C., Domoney, C., & Brites, C. (2019). Variation in pea (*Pisum sativum* L.) seed quality traits defined by physicochemical functional properties. *Foods, 8*(11), 570. https://doi.org/10.3390/foods8110570

Sharma, A., & Gupta, M. (2023). Characterization of physicochemical, functional and antioxidant properties of western Himalayan black pea. *Journal of Agriculture and Food Research, 13*, Article 100607. https://doi.org/10.1016/j.jafr.2023.100607

Shen, Y., Hong, S., & Li, Y. (2022). Pea protein composition, functionality, modification, and food applications: A review. *Advances in Food and Nutrition Research, 101*, 71–127. https://doi.org/10.1016/bs.afnr.2022.02.002

Shi, J., Zeng, K., Guo, D., Wang, P., Zhang, S., Ren, F., & Liu, S. (2023). Insights into the relation between multi-scale structure and in-vitro digestibility of type 3 resistant starches prepared from wrinkled pea starches. *Food Hydrocolloids, 144*, Article 109056. https://doi.org/10.1016/j.foodhyd.2023.109056

Sun, C., Ge, J., He, J., Gan, R., & Fang, Y. (2021). Processing, quality, safety, and acceptance of meat analogue products. *Engineering, 7*(5), 674–678. https://doi.org/10.1016/j.eng.2020.10.011

Sun, G., Ni, P., Lam, E., Hrapovic, S., Bing, D., Yu, B., & Ai, Y. (2023). Exploring the functional attributes and in vitro starch and protein digestibility of pea flours having a wide range of amylose content. *Food Chemistry, 405*(Part B), Article 134938. https://doi.org/10.1016/j.foodchem.2022.134938

Tanger, C., Engel, J., & Kulozik, U. (2020). Influence of extraction conditions on the conformational alteration of pea protein extracted from pea flour. *Food Hydrocolloids, 107*, Article 105949. https://doi.org/10.1016/j.foodhyd.2020.105949

Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., … Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ, 2014*(1), Article e453. https://doi.org/10.7717/peerj.453

Vreeke, G. J. C., Meijers, M. G. J., Vincken, J. P., & Wierenga, P. A. (2023). Towards absolute quantification of protein genetic variants in Pisum sativum extracts. *Analytical Biochemistry, 665*(15), Article 115048. https://doi.org/10.1016/j.ab.2023.115048

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., … Zong, X. (2022). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nature Genetics, 54*(10), 1553–1563. https://doi.org/10.1038/s41588-022-01172-2

Zhang, L., Li, Q., Zhang, W., Bakalis, S., Luo, Y., & Lametsch, R. (2024). Different source of commercial soy protein isolates: Structural, compositional, and physicochemical characteristics in relation to protein functionalities. *Food Chemistry, 433*, Article 137315. https://doi.org/10.1016/j.foodchem.2023.137315

Zhong, L., Fang, Z., Wahlqvist, M. L., Wu, G., Hodgson, J. M., & Johnson, S. K. (2018). Seed coats of pulses as a food ingredient: Characterization, processing, and applications. *Trends in Food Science and Technology, 80*, 35–42. Elsevier Ltd https://doi.org/10.1016/j.tifs.2018.07.021.