

## NRC Publications Archive Archives des publications du CNRC

### Implementation of machine learning models to predict functionality of pea flour from its composition

Nickerson, Colten N.; Hogarth, Sarah; Stone, Andrea K.; Arganosa, Gene C.; Bhowmik, Pankaj; Warkentin, Tom D.; Ubbens, Jordan; Nickerson, Michael T.

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

#### **Publisher's version / Version de l'éditeur:**

<https://doi.org/10.1002/cche.70072>

*Cereal Chemistry*, pp. 1-15, 2026-05-06

#### **NRC Publications Archive Record / Notice des Archives des publications du CNRC :**

<https://nrc-publications.canada.ca/eng/view/object/?id=169bd387-2e60-4470-a499-af205e4cf3e2>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=169bd387-2e60-4470-a499-af205e4cf3e2>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.





**Questions?** Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

**Vous avez des questions?** Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

## RESEARCH ARTICLE OPEN ACCESS

# Implementation of Machine Learning Models to Predict Functionality of Pea Flour From Its Composition

Colten N. Nickerson<sup>1</sup> | Sarah Hogarth<sup>2</sup> | Andrea K. Stone<sup>2</sup> | Gene C. Arganosa<sup>3</sup>  | Pankaj Bhowmik<sup>1</sup>  | Tom D. Warkentin<sup>3</sup>  | Jordan Ubbens<sup>1</sup> | Michael T. Nickerson<sup>2</sup> 

<sup>1</sup>Aquatic and Crop Resource Development, National Research Council Canada, Saskatoon, Saskatchewan, Canada | <sup>2</sup>Department of Food and Bioproduct Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada | <sup>3</sup>Department of Plant Sciences, Crop Development Centre, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

**Correspondence:** Michael T. Nickerson ([Michael.Nickerson@usask.ca](mailto:Michael.Nickerson@usask.ca))

**Received:** 24 October 2025 | **Revised:** 23 March 2026 | **Accepted:** 21 April 2026

**Funding:** Pulse Science Cluster, Grant/Award Number: ASC-2023-15B; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: CREATE-CAPTURE

**Keywords:** Gaussian process regression | multi-layer perceptron | neural network | solubility | support vector regression | XGBoost

## ABSTRACT

**Background and Objectives:** The goal of this research was to examine the relationship between the composition and functionality of pea flour using the following machine learning algorithms: linear regression, partial least squares regression (PLSR), Gaussian process regression (GPR), support vector regression, gradient-boosted decision trees, and a standard feed-forward neural network.

**Findings:** In general, linear models outperformed non-linear models. PLSR provided best fits for prediction of emulsion stability, oil holding capacity, foam stability and foam capacity; but was less effective for solubility and water holding capacity, which were best described by the GPR model. Variable Importance in Projection scores, calculated for each PLSR model, showed that protein and acid detergent fiber were both highly influential in predicting foaming capacity (1.52 and 1.55), foaming stability (1.30 and 1.54), oil holding capacity (1.64 and 1.50), and water holding capacity (1.56 and 1.53). Protein was also highly important in predicting solubility (1.80), alongside starch (1.60) whereas lipid was highly predictive (2.02) for emulsion stability.

**Conclusion:** Application of machine learning models was successful in relating compositional features of pea flour to functionality.

**Significance and Novelty:** Using machine learning to predict the functional behavior of pea will aid both breeders and product developers in ingredient selection.

## 1 | Introduction

Over the last decade, significant advances and uptake of alternative protein ingredients by the food industry have occurred to reduce or replace the use of soybean, wheat, and animal-derived proteins. Market shifts are being driven by issues surrounding cost, environmental sustainability,

changing regulatory frameworks, changing consumer demographics and health/nutrition (Cordoba et al. 2025). Pulse proteins, particularly those derived from pea, are attractive due to their low cost, abundant supply, and nutritional/functional properties. Challenges exist related to ingredient performance, as food companies aim to integrate pea ingredients into various food market segments. Suppliers tend to

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 His Majesty the King in Right of Canada and The Author(s). *Cereal Chemistry* published by Wiley Periodicals LLC on behalf of Cereals & Grains Association. Reproduced with the permission of the Minister of Innovation, Science, and Economic Development.

see variability in the feedstock composition associated with genetic and environmental effects that can have a direct impact on functionality of the ingredient, and thus their performance in food products. Knowledge surrounding the strength of correlations tying seed composition to functionality is limited. Integration with machine learning driven models to better understand and predict this relationship could lead to advances in breeding programs and ingredient performance. Artificial neural networks can predict pea seed yield with high accuracy, significantly outperforming traditional linear models by effectively utilizing meteorological, agronomic, and phytophysical data (Hara et al. 2022). Kircali Ata et al. (2023) reported the hardness and chewiness prediction of plant-based meat analogs produced by extrusion could best be described by their relationship with the proximate composition of the ingredients using a regularized line Ridge model.

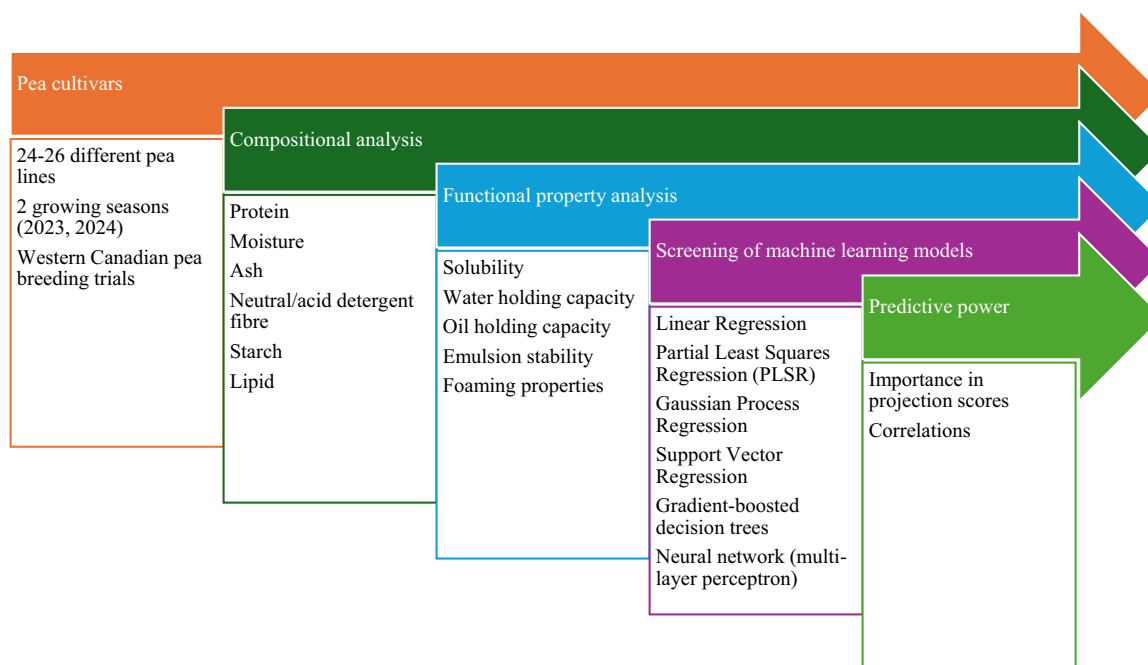
Machine learning models were used to predict the protein content of peas by analyzing their spectral data, enabling rapid and non-destructive assessment for applications in biology and food science (Xie et al. 2024). Machine learning methods can capture complex, non-linear relationships among independent and dependent variables (Lie-Piang et al. 2023). Many machine learning algorithms are available for tabular data, however, few studies have investigated their utility for mapping ingredient functionality to performance. Screening and selection of the best model for this purpose is an essential activity of this research, including assessing the feasibility and predictive ability of candidate models.

Lie-Piang et al. (2023) used machine learning (e.g., spline regression, random forest, and neural networks) to predict and quantify functional attributes of a wide range of ingredient formulations. The authors considered functional attributes of commercial isolates (e.g., pea protein isolate, pea fiber isolate, pea starch isolate, and lupin protein isolates) and mildly refined

ingredients (e.g., flour, air classified protein and starch fractions for both pea and lupin), which included emulsion stability, foaming capacity, gel stiffness, and final viscosities, and mapped those to the formulation composition. The authors reported that different machine learning models needed to be individualized for each functionality and crop.

Dahl et al. (2025) predicted the rheological properties from compositional data arising from food biopolymer mixtures using machine learning. Formulations involving pea and faba bean protein isolates/concentrates (14%–43%) were blended with various polysaccharides (0%–13%) including maize starch, pectin, cellulose and carrageenan, with final moisture levels ranging between 40% and 72%. These levels were selected to be indicative of high moisture extrusion processing/formulation conditions. Data sets were evaluated using cluster analysis to identify patterns in the data sets and variable importance, prior to applying machine learning models. The authors reported that when samples were within the linear viscoelastic regime, a single-output random forest model was most effective at relating compositional data to its rheological data. For those exhibiting a non-linear response, a multi-output random forest regression was more effective. Neural network models were also applied to predict spray drying conditions for encapsulation of probiotics using pea proteins (Sharma et al. 2024). For the most part, in the food industry, deep learning algorithms have focused on image analysis as it relates to processing, quality and safety (Zhu et al. 2021). As an example, Wu et al. (2023) studied the impact of residence time and texture prediction of pea protein extrusion based on image analysis.

The overall goal of this research was to screen various machine learning models for their predictive power in relating pea seed/flour composition to a range of functional properties important to the food industry (Figure 1). The impact of pea cultivar and growing season are included to improve the robustness of the dataset.



**FIGURE 1** | Experimental design for screening various machine learning models to predict the functionality of pea flour from their composition, for varying pea cultivars and growing seasons from Western Canada.

## 2 | Materials and Methods

### 2.1 | Materials

Pea breeding lines from the 2023 and 2024 western Canadian pea Co-op Test, with three replicates each year, were evaluated for protein composition and functionality traits. The 2023 lines consisted of 12 from pea Co-op Test B grown at Sutherland (Saskatoon) location, that is, CDC Spectrum, AAC Lacombe, AAC Profit, CDC 7088-F1-21-1Y, CDC Limerick, CDC Forest, CDC 6549-4, CDC 7084-F1-7-62, CDC 7090-F1-7-91, CDC 7088-F1-21-1G, CDC 6897-6, and CDC 6857-1, and 12 from pea Co-op Test A also grown at Sutherland (Saskatoon) location, that is, CDC 6461-1, CDC 6482-4, CDC 6308-3, CDC 6471-2, CDC 6844-9, CDC 6736-2, CDC 6680-1, CDC 6716-14, CDC 6635-2, CDC 6735-2, CDC 6735-4, and CDC 6862-1. The 2024 lines consisted of 7 from pea Co-op Test B, that is, CDC 7084-F1-7-62, CDC 6897-6, CDC 6948-4, CDC 7078-2, CDC 7088-F1-21-1Y, CDC 6999-2 and CDC 7101-13, and 19 from pea Co-op Test A, that is, CDC Limerick, CDC Forest, CDC Spectrum, AAC Profit, CDC 6844-9, CDC 6736-2, CDC 6680-1, CDC 6716-14, CDC 7007-15, CDC 7144-2, CDC 7100-2, CDC 6959-7, CDC 7062-5, CDC 7150-5, CDC 6933-5, CDC 6921-4, CDC 7113-5, CDC 7040-3, and CDC 7027-3. The whole pea seeds were milled into flour using a benchtop mill (Wondermill, Garin Mill model: WM2000, Pocatello, ID). All milled flours were stored in sealable plastic (low density polyethylene, 0.07 mm thickness) bags at 4°C until further analysis. All water used was sourced from a MilliQ™ water purification system (Millipore Corp, Milford, MA, USA).

### 2.2 | Compositional Analysis by Near Infrared Spectroscopy

Compositional analysis (protein, moisture, ash, neutral/acid detergent fiber, starch and crude lipid (i.e., ether extract)) were determined using a FOSS NIR DS2500 Near Infrared spectrophotometer (FOSS North America, Eden Prairie, MN, USA), where reflectance of the pea samples was scanned from 400 to 2498 nm in 0.5 nm increments (Arganosa et al. 2006).

### 2.3 | Flour Functionality

#### 2.3.1 | Protein Solubility

Protein solubility (%) was determined by dispersing 0.200 g of flour in 20 mL water. The samples were adjusted to pH 7.0 using 0.1 N NaOH and stirred (300 rpm) for 1 h at room temperature (22°C). A corresponding sample was prepared by dispersing 0.200 g of flour in 20 mL of 1 N NaOH (to completely solubilize the sample) and allowing it to stir as previously described. The samples were then centrifuged at 7000g for 15 min at room temperature (Sorvall ST8 Centrifuge, Thermo Fisher Scientific Inc., Waltman, MA, USA) and stored overnight at 4°C. The protein content of the supernatant was determined using the modified Lowry Method (Thermo Scientific Modified Lowry Protein Assay Kit, LSG23240) according to Lo et al. (2022). A dilution (0.3 mL supernatant in 10 mL water) was used to ensure the absorbance readings were within the bovine serum albumin standard curve. Percent solubility was calculated by dividing the protein content of the

supernatant at pH 7 by the protein content of the sample in 1 N NaOH (×100%).

#### 2.3.2 | Water and Oil Holding Capacity

Water holding capacity (WHC) was measured by weighing 1.0 g flour and 10 g water into a 50 mL centrifuge tube according to Stone et al. (2015). The mixture was vortexed at maximum speed 10 (Analog Vortex Mixer, VWR International, Mississauga, ON, Canada) for 10 s every 5 min during a 30 min period and then centrifuged at 1000g for 15 min (Sorvall ST8 Centrifuge, Thermo Fisher Scientific Inc., Waltman, MA, USA). The supernatant was subsequently drained from the centrifuge tube and the mass of the resulting pellet was weighed. The amount of water absorbed by the flour was determined based on the difference in pellet weight and initial flour weight (reported as g water/g flour). The same method was used to determine oil holding capacity (OHC; g oil/g flour), using 10 g canola oil in place of water.

$$\text{WHC or OHC} = \frac{[\text{pellet weight (g)} - \text{initial flour weight (g)}]}{\text{initial flour weight (g)}} \quad (1)$$

#### 2.3.3 | Foaming

Foaming capacity (FC) and foam stability (FS) were measured by preparing a 1% flour sample in water according to Stone et al. (2015) with slight modifications. The solution was stirred for 20 min using a magnetic stir plate followed by pH adjustment to pH 7.0. After stirring for 1 h at pH 7.0, 15 mL of the flour solution was transferred into a 400 mL beaker (inner diameter = 69 mm; height = 127 mm). Using a homogenizer (Polytron PT2100, Kinematica AG, Lucerne, Switzerland) with a saw tooth generating probe positioned slightly below the air-water interface, the solution was homogenized at 13,000 rpm for 2 min to generate foam. Immediately following, the foam was poured into a 50 mL graduated cylinder and the foam volume (FV1) was recorded. After 30 min the foam volume (FV2) was again recorded to measure the FS.

$$\%FC = \frac{\text{FV1 (mL)}}{15 \text{ mL}} \times 100 \quad (2)$$

$$\%FS = \frac{\text{FV2 (mL)}}{\text{FV1 (mL)}} \times 100. \quad (3)$$

#### 2.3.4 | Emulsion Stability (ES)

ES by creaming was determined by preparing a 1% protein sample in water and adjusting the solution pH to 7.0 over 1 h at room temperature according to Stone et al. (2015). Emulsions were prepared by homogenizing 5 mL (V1) of the prepared 1% protein solution with 5 g of canola oil in 50 mL centrifuge tubes. Mixtures were homogenized for 2 min at 13,000 rpm using a Polytron homogenizer (as described for foaming). The emulsion was immediately transferred into a 10 mL graduated glass cylinder (inner diameter = 10.80 mm; height = 100.24 mm) after preparation. The ES was determined by observing the separation of the serum layer after 30 min of storage time at room temperature (V2). The emulsion stability was calculated using the following calculation:

$$\%ES = \frac{(V1 - V2)}{V1} \times 100. \quad (4)$$

**TABLE 1** | Compositional analysis by near infrared spectroscopy of protein, starch, neutral detergent fiber, acid detergent fiber, lipid, and ash (% dry basis) for the 2023 and 2024 growing seasons at Saskatoon, Canada. Data represent the mean and standard deviation of triplicate plots.

<b>Cultivar</b>	<b>Protein (%)</b>	<b>Starch (%)</b>	<b>Neutral detergent fiber (%)</b>	<b>Acid detergent fiber (%)</b>	<b>Lipid (%)</b>	<b>Ash (%)</b>
2023 Growing season						
CDC Spectrum	22.3 ± 0.9	49.4 ± 2.7	14.9 ± 0.7	11.6 ± 1.6	0.8 ± 0.2	2.4 ± 0.2
AAC Lacombe	20.7 ± 0.5	50.6 ± 1.1	14.2 ± 1.1	10.6 ± 0.6	1.2 ± 0.1	2.4 ± 0.0
CDC Limerick	22.5 ± 0.9	48.6 ± 1.6	14.0 ± 1.1	10.6 ± 0.6	1.2 ± 0.1	2.4 ± 0.0
CDC Forest	21.5 ± 0.6	49.4 ± 1.2	14.1 ± 1.1	11.6 ± 0.3	1.1 ± 0.3	2.3 ± 0.1
AAC Profit	22.9 ± 0.4	47.6 ± 1.4	13.2 ± 0.5	11.1 ± 0.3	0.6 ± 0.4	2.3 ± 0.1
CDC 6549-4	23.2 ± 1.2	47.4 ± 0.4	13.8 ± 0.7	11.0 ± 0.2	1.1 ± 0.1	2.5 ± 0.1
CDC 7084-F1-7-62	22.1 ± 0.9	47.3 ± 1.0	16.5 ± 0.2	12.2 ± 0.3	1.3 ± 0.2	2.6 ± 0.0
CDC 7090-F1-7-91	23.4 ± 0.3	47.2 ± 0.9	13.4 ± 1.8	11.7 ± 1.4	1.0 ± 0.3	2.2 ± 0.0
CDC 7088-F1-21-1Y	22.2 ± 1.1	48.7 ± 0.8	14.4 ± 0.3	12.2 ± 0.2	0.3 ± 0.3	2.3 ± 0.1
CDC 7088-F1-21-1G	22.8 ± 1.2	48.7 ± 1.5	13.2 ± 1.0	11.4 ± 1.0	1.0 ± 0.3	2.3 ± 0.0
CDC 6897-6	21.3 ± 1.2	48.5 ± 1.8	14.5 ± 0.6	11.5 ± 1.0	1.4 ± 0.4	2.3 ± 0.1
CDC 6857-1	20.9 ± 0.7	49.5 ± 1.8	14.1 ± 1.1	11.4 ± 1.0	0.9 ± 0.1	2.3 ± 0.1
CDC 6461-1	22.0 ± 0.6	50.7 ± 0.8	13.5 ± 0.9	11.6 ± 0.9	1.8 ± 0.2	1.7 ± 0.1
CDC 6482-4	21.3 ± 0.1	53.2 ± 0.3	12.9 ± 0.3	10.8 ± 0.2	2.1 ± 0.1	1.8 ± 0.1
CDC 6308-3	20.4 ± 0.7	51.3 ± 1.3	13.3 ± 0.5	12.3 ± 0.6	1.8 ± 0.1	1.7 ± 0.2
CDC 6471-2	22.3 ± 0.9	49.5 ± 1.1	13.4 ± 0.7	12.6 ± 0.4	1.8 ± 0.1	1.7 ± 0.2
CDC 6844-9	22.3 ± 0.3	50.2 ± 1.3	13.3 ± 1.2	11.7 ± 0.7	1.3 ± 0.2	1.7 ± 0.2
CDC 6736-2	21.6 ± 0.2	50.1 ± 1.3	13.1 ± 0.7	11.5 ± 0.7	1.6 ± 0.1	1.6 ± 0.0
CDC 6680-1	21.4 ± 0.2	49.0 ± 3.4	14.1 ± 1.4	13.0 ± 1.0	1.3 ± 0.2	1.6 ± 0.2
CDC 6716-14	24.2 ± 0.5	49.2 ± 3.4	14.1 ± 1.4	13.0 ± 1.0	1.3 ± 0.2	1.6 ± 0.2
CDC 6635-2	20.3 ± 0.3	52.0 ± 0.36	13.1 ± 0.3	11.5 ± 0.4	1.8 ± 0.2	1.6 ± 0.1
CDC 6735-2	20.9 ± 1.2	51.1 ± 1.6	14.7 ± 0.1	11.9 ± 0.8	1.7 ± 0.3	1.8 ± 0.2
CDC 6735-4	20.9 ± 0.7	48.4 ± 1.0	16.4 ± 0.2	12.6 ± 0.7	1.9 ± 0.3	1.9 ± 0.1
CDC 6862-1	21.3 ± 1.0	50.9 ± 1.8	14.4 ± 0.6	12.1 ± 0.8	1.6 ± 0.2	1.6 ± 0.0
Mean (2023)	21.9 ± 1.2	49.5 ± 1.9	14.0 ± 1.2	11.7 ± 0.9	1.3 ± 0.5	2.0 ± 0.4
Range (2023)	20.3–23.4	47.2–53.2	12.9–16.5	10.6–13.0	0.6–2.1	1.6–2.6
2024 Growing season						
CDC Spectrum	27.8 ± 0.5	49.9 ± 0.5	15.1 ± 1.3	9.4 ± 0.3	1.4 ± 0.3	2.4 ± 0.1
AAC Profit	27.7 ± 0.9	48.1 ± 1.5	15.0 ± 2.0	9.6 ± 0.6	1.4 ± 0.2	2.3 ± 0.1
CDC Limerick	29.3 ± 1.2	50.7 ± 0.6	12.8 ± 0.2	8.8 ± 0.3	1.9 ± 0.2	2.3 ± 0.1
CDC Forest	27.3 ± 1.3	52.4 ± 0.8	13.8 ± 0.7	9.2 ± 0.9	2.3 ± 0.2	2.4 ± 0.1
CDC 6844-9	26.3 ± 0.9	50.8 ± 0.9	14.2 ± 1.0	8.9 ± 0.6	1.7 ± 0.3	2.1 ± 0.1
CDC 6736-2	26.6 ± 1.4	51.2 ± 0.9	13.4 ± 0.8	8.6 ± 0.6	2.1 ± 0.6	2.0 ± 0.1
CDC 6680-1	25.3 ± 1.1	51.9 ± 0.8	14.4 ± 0.3	9.1 ± 0.3	1.3 ± 0.3	2.2 ± 0.0
CDC 6716-14	26.7 ± 1.4	49.9 ± 1.1	14.0 ± 0.5	8.3 ± 0.5	1.7 ± 0.4	2.2 ± 0.1
CDC 7007-15	27.5 ± 0.5	48.5 ± 0.7	14.5 ± 1.4	9.3 ± 0.8	1.4 ± 0.5	2.1 ± 0.1
CDC 7144-2	27.2 ± 0.9	51.3 ± 0.8	13.3 ± 0.5	8.4 ± 0.5	1.7 ± 0.3	2.1 ± 0.1
CDC 7100-2	25.7 ± 0.8	52.6 ± 0.7	14.1 ± 0.9	8.4 ± 0.5	1.7 ± 0.3	2.2 ± 0.0
CDC 6959-7	27.0 ± 0.9	51.3 ± 0.9	14.2 ± 0.5	9.6 ± 0.3	1.4 ± 0.1	2.2 ± 0.0
CDC 7062-5	26.3 ± 1.0	49.8 ± 0.8	14.8 ± 0.4	9.0 ± 0.6	1.8 ± 0.2	2.2 ± 0.0
CDC 7150-5	25.1 ± 0.5	51.6 ± 1.9	15.0 ± 0.6	9.1 ± 0.9	1.7 ± 0.2	2.2 ± 0.1
CDC 6933-5	28.3 ± 0.8	50.9 ± 1.4	12.8 ± 0.5	8.0 ± 0.3	1.7 ± 0.2	2.1 ± 0.0
CDC 6921-4	27.8 ± 1.1	49.7 ± 0.6	13.2 ± 0.8	8.6 ± 0.6	1.4 ± 0.4	2.2 ± 0.0

(Continues)

TABLE 1 | (Continued)

Cultivar	Protein (%)	Starch (%)	Neutral detergent	Acid detergent	Lipid (%)	Ash (%)
			fiber (%)	fiber (%)		
CDC 7113-5	27.5 ± 0.8	50.0 ± 1.2	13.6 ± 0.2	8.7 ± 0.6	1.6 ± 0.4	2.2 ± 0.0
CDC 7040-3	27.5 ± 1.2	49.8 ± 0.7	13.7 ± 0.4	8.6 ± 0.6	1.4 ± 0.2	2.1 ± 0.0
CDC 7027-3	25.9 ± 1.4	53.0 ± 1.2	14.6 ± 0.4	8.8 ± 0.6	1.7 ± 0.2	2.1 ± 0.1
CDC 7084-F1-7-62	28.1 ± 0.6	50.6 ± 1.1	15.2 ± 0.3	10.1 ± 0.7	2.3 ± 0.5	2.5 ± 0.1
CDC 7088-F1-21-1Y	28.3 ± 0.4	47.5 ± 0.8	14.7 ± 0.8	10.1 ± 0.4	1.1 ± 0.3	2.3 ± 0.0
CDC 6897-6	26.3 ± 1.4	51.7 ± 1.0	12.6 ± 0.4	8.9 ± 0.6	2.4 ± 0.3	2.2 ± 0.1
CDC 6948-4	26.8 ± 1.3	52.6 ± 0.8	13.0 ± 0.5	9.2 ± 0.2	1.8 ± 0.3	2.1 ± 0.1
CDC 7078-2	26.5 ± 1.3	50.8 ± 1.6	14.0 ± 1.4	9.6 ± 0.8	2.1 ± 0.6	2.3 ± 0.1
CDC 6999-2	27.6 ± 0.6	49.3 ± 1.0	14.2 ± 0.4	9.0 ± 0.6	1.4 ± 0.2	2.2 ± 0.1
CDC 7101-13	28.6 ± 1.0	49.4 ± 1.4	13.0 ± 0.4	8.8 ± 0.3	1.4 ± 0.2	2.1 ± 0.1
Mean (2024)	27.1 ± 1.3	50.6 ± 1.6	14.0 ± 0.9	9.0 ± 0.7	1.7 ± 0.4	2.2 ± 0.1
Range (2024)	25.1–29.3	47.5–52.6	12.8–15.2	8.0–10.1	1.1–2.3	2.1–2.4

## 2.4 | Machine Learning Models

In total, six candidate models were selected, representing a diverse set of linear and non-linear approaches. For linear models, Linear Regression (LR) and Partial Least Squares Regression (PLSR) were evaluated. For non-linear models, Gaussian Process Regression (GPR), Support Vector Regression (SVR), gradient-boosted decision trees, and a standard feed-forward neural network (multi-layer perceptron, MLP) were evaluated.

PLSR is a linear method which excels in contexts where there is a high degree of multicollinearity among the predictors, which makes it a robust and common choice in many application areas such as chemometrics. PLSR learns a projection of the features into a reduced latent space which is maximally correlated with the dependent variable (or variables). This lends to not only robustness against correlated features, but also interpretable predictions by inspecting the weights of individual components.

GPR is a Bayesian regression method which assumes a Gaussian prior distribution over the data generating process. Given a kernel function  $K$  and data  $x$ , GPR models condition the class of functions defined by  $K$  based on the observations, with the assumption that they were sampled from a multivariate Gaussian distribution with  $K$  defining its covariance. GPR has the advantage of giving not only point predictions but the full posterior distribution over predictive functions, although this comes at the cost of high computational complexity. For the kernel function, we evaluated the RBF (Gaussian) kernel, the dot product kernel, and the rational quadratic kernel.

SVR generalizes support vector machines, a common model for classification, to continuous regression targets. SVR models learn to fit a function  $f(x)$  such that the training points lie within a margin around it and only points which exceed this threshold contribute to the loss during fitting. This function can be a linear function (i.e.  $f(x) = wx + b$ ), or non-linear using a kernel function. We test the linear kernel, as well as the polynomial kernel, the RBF kernel, and the sigmoid kernel.

Gradient boosted decision trees are an ensemble method which combine multiple decision trees, which act as weak learners. Each

decision tree is fit to minimize the residual error produced by the previous tree and the ensemble is heavily regularized to mitigate overfitting. There are several implementations of gradient-boosted decision trees available—in this work we use the popular XGBoost implementation (Chen and Guestrin 2016).

Multi-layer perceptrons (or feed-forward neural networks) are a classical example of neural networks. Inputs are successively transformed by repeated linear projections, each followed by a non-linear transformation. The output of the final transformation is compared to the ground truth using a loss function (here we use mean squared error), and the parameters of the model are updated by backpropagation (against the gradient of the loss with respect to each parameter). For the MLP model, dropout stochastic regularization (Srivastava et al. 2014) was used between hidden layers. The architecture of the MLP as well as the optimizer and other training hyperparameters were determined via hyperparameter search (Section 3.3).

## 3 | Results and Discussions

### 3.1 | Compositional Analysis by Near Infrared Spectroscopy

Compositional analysis of 24 and 26 pea lines from the 2023 and 2024 growing season are reported in Table 1. Overall, mean protein levels in pea breeding lines were lower in 2023 (21.9%) than in 2024 (27.1%). In contrast, starch (~50%), neutral detergent fiber (~14%) and ash levels were similar (~2.1%) across years. Mean acid detergent fiber was greater in 2023 (11.7%) relative to 2024 (9.0%); whereas lipid levels were higher in 2024 (1.7%) than 2023 (1.3%). Values reflect mean values of all lines, where some variation among the pea lines within each year was evident (Table 1). Protein levels are within the range of other reports for peas grown in Saskatchewan (Canada). For instance, Stone et al. (2019), reported pea flour (CDC Meadow) grown in Saskatchewan to have approximately 24.5% protein; Hood-Niefer et al. (2012) reported variation in pea flour protein levels (24.2%–27.5%) among 10 pea lines; and Nosworthy et al. (2021) reported protein levels ranging

**TABLE 2** | Functional properties of pea flour from seeds from the 2023 and 2024 growing seasons at Saskatoon, Canada. Data represent the mean and standard deviation of triplicate plots.

<b>Cultivar</b>	<b>Solubility (%)</b>	<b>Water holding capacity (g/g)</b>	<b>Oil holding capacity (g/g)</b>	<b>Foaming capacity (%)</b>	<b>Foam stability (%)</b>	<b>Emulsion stability (%)</b>
2023 Growing season						
CDC Spectrum	68.7 ± 2.3	1.28 ± 0.05	0.88 ± 0.03	85.6 ± 1.9	90.9 ± 2.1	84.3 ± 0.6
AAC Lacombe	76.2 ± 4.2	1.26 ± 0.05	0.87 ± 0.01	98.9 ± 1.9	85.4 ± 2.3	79.7 ± 0.6
CDC Limerick	78.5 ± 2.8	1.39 ± 0.02	0.88 ± 0.03	98.9 ± 5.1	92.4 ± 6.8	82.7 ± 1.2
CDC Forest	77.4 ± 1.9	1.33 ± 0.03	0.87 ± 0.01	110.0 ± 3.3	85.9 ± 3.5	80.3 ± 0.6
AAC Profit	86.3 ± 1.1	1.38 ± 0.05	0.91 ± 0.04	103.3 ± 5.8	84.9 ± 2.4	88.0 ± 1.7
CDC 6549-4	78.8 ± 1.5	1.37 ± 0.05	0.90 ± 0.02	115.6 ± 7.7	89.4 ± 2.1	83.7 ± 1.5
CDC 7084-F1-7-62	84.3 ± 0.7	1.37 ± 0.08	0.93 ± 0.01	134.4 ± 9.6	86.8 ± 1.2	83.7 ± 2.3
CDC 7090-F1-7-91	79.6 ± 2.0	1.42 ± 0.06	0.93 ± 0.04	117.8 ± 13.5	88.6 ± 1.2	81.3 ± 0.6
CDC 7088-F1-21-1Y	80.6 ± 0.1	1.31 ± 0.07	0.97 ± 0.01	123.3 ± 3.3	88.2 ± 3.4	93.0 ± 1.0
CDC 7088-F1-21-1G	78.2 ± 0.1	1.33 ± 0.01	0.92 ± 0.03	148.9 ± 1.9	89.6 ± 1.2	91.0 ± 1.7
CDC 6897-6	79.0 ± 3.2	1.33 ± 0.06	0.91 ± 0.01	142.2 ± 12.6	87.4 ± 2.0	87.0 ± 1.7
CDC 6857-1	77.4 ± 3.6	1.27 ± 0.04	0.91 ± 0.02	131.1 ± 13.9	87.2 ± 1.3	89.3 ± 0.6
CDC 6461-1	80.7 ± 1.9	1.35 ± 0.02	0.86 ± 0.02	150.0 ± 16.7	87.3 ± 2.0	83.7 ± 0.6
CDC 6482-4	74.4 ± 1.8	1.37 ± 0.03	0.88 ± 0.04	190.0 ± 9.77	90.0 ± 1.6	82.0 ± 1.0
CDC 6308-3	73.3 ± 1.7	1.34 ± 0.04	0.90 ± 0.01	186.7 ± 27.3	84.7 ± 5.1	82.0 ± 0.0
CDC 6471-2	79.5 ± 4.4	1.34 ± 0.04	0.86 ± 0.01	175.6 ± 34.7	88.2 ± 3.3	89.3 ± 1.2
CDC 6844-9	80.9 ± 4.9	1.33 ± 0.03	0.89 ± 0.01	105.6 ± 1.9	92.6 ± 2.0	92.3 ± 0.6
CDC 6736-2	79.7 ± 2.6	1.30 ± 0.05	0.89 ± 0.03	123.3 ± 3.3	84.7 ± 1.6	86.0 ± 2.0
CDC 6680-1	82.7 ± 3.2	1.51 ± 0.04	0.89 ± 0.01	110.0 ± 3.3	89.9 ± 1.5	84.7 ± 3.1
CDC 6716-14	79.7 ± 1.3	1.30 ± 0.01	0.86 ± 0.02	137.8 ± 10.2	89.4 ± 3.1	85.7 ± 1.5
CDC 6635-2	81.7 ± 1.5	1.35 ± 0.01	0.88 ± 0.02	108.9 ± 1.9	86.7 ± 1.7	80.0 ± 2.0
CDC 6735-2	77.6 ± 2.7	1.41 ± 0.04	0.89 ± 0.02	131.1 ± 5.1	89.8 ± 2.8	81.0 ± 1.7
CDC 6735-4	85.5 ± 0.5	1.32 ± 0.02	0.90 ± 0.04	124.4 ± 5.1	88.4 ± 1.7	82.0 ± 1.0
CDC 6862-1	79.1 ± 2.9	1.34 ± 0.05	0.89 ± 0.00	122.2 ± 3.8	88.2 ± 1.8	90.0 ± 1.0
Mean	76.2 ± 3.8	1.35 ± 0.05	0.89 ± 0.03	128.2 ± 27.0	88.2 ± 2.2	85.1 ± 4.0
Range	68.7–86.3	1.26–1.51	0.86–0.97	85.6–190.0	84.7–92.6	79.7–93.0
2024 Growing season						
CDC Spectrum	86.0 ± 2.0	1.41 ± 0.03	1.17 ± 0.02	92.9 ± 3.4	84.4 ± 7.6	84.0 ± 1.7
AAC Profit	86.8 ± 1.3	1.43 ± 0.07	1.35 ± 0.05	105.6 ± 11.7	86.1 ± 3.7	82.1 ± 1.2
CDC Limerick	83.1 ± 3.1	1.16 ± 0.04	0.98 ± 0.00	74.4 ± 5.1	80.1 ± 9.3	84.5 ± 3.2
CDC Forest	77.8 ± 1.8	1.21 ± 0.01	0.95 ± 0.04	90.4 ± 9.1	76.4 ± 5.1	83.5 ± 1.2
CDC 6844-9	76.8 ± 0.3	1.20 ± 0.07	1.01 ± 0.04	108.7 ± 4.4	78.6 ± 8.2	82.8 ± 2.0
CDC 6736-2	77.5 ± 3.1	1.21 ± 0.06	1.01 ± 0.04	70.2 ± 12.5	78.8 ± 10.8	85.5 ± 1.2
CDC 6680-1	77.8 ± 0.9	1.35 ± 0.05	1.05 ± 0.05	110.0 ± 5.8	79.7 ± 4.2	86.1 ± 2.1
CDC 6716-14	70.5 ± 0.3	1.20 ± 0.04	1.06 ± 0.05	110.0 ± 3.3	78.8 ± 5.8	85.5 ± 1.2
CDC 7007-15	76.5 ± 2.6	1.16 ± 0.05	1.03 ± 0.01	94.4 ± 4.1	77.8 ± 6.9	83.8 ± 1.0
CDC 7144-2	85.6 ± 3.4	1.13 ± 0.02	0.97 ± 0.03	102.2 ± 10.2	75.4 ± 2.6	83.8 ± 3.6
CDC 7100-2	79.7 ± 1.2	1.32 ± 0.02	0.99 ± 0.01	104.9 ± 7.3	76.5 ± 4.7	84.1 ± 1.2
CDC 6959-7	83.6 ± 1.5	1.23 ± 0.05	1.03 ± 0.05	101.1 ± 3.8	79.7 ± 4.5	83.6 ± 3.1
CDC 7062-5	78.9 ± 1.8	1.09 ± 0.06	1.11 ± 0.05	110.4 ± 5.4	85.6 ± 1.8	83.5 ± 0.9
CDC 7150-5	79.5 ± 1.4	1.21 ± 0.06	1.18 ± 0.05	101.3 ± 12.7	84.3 ± 3.7	82.7 ± 2.3
CDC 6933-5	79.3 ± 3.0	1.18 ± 0.05	1.05 ± 0.02	109.6 ± 7.1	84.2 ± 4.2	86.7 ± 1.2
CDC 6921-4	79.4 ± 2.9	1.22 ± 0.05	1.07 ± 0.02	68.9 ± 7.7	75.6 ± 2.5	86.7 ± 1.2

(Continues)

TABLE 2 | (Continued)

Cultivar	Solubility (%)	Water holding capacity (g/g)	Oil holding capacity (g/g)	Foaming capacity (%)	Foam stability (%)	Emulsion stability (%)
CDC 7113-5	79.4 ± 1.7	1.15 ± 0.09	1.15 ± 0.07	89.3 ± 2.3	79.5 ± 6.7	86.0 ± 1.7
CDC 7040-3	76.1 ± 1.2	1.19 ± 0.05	1.24 ± 0.07	101.9 ± 9.5	84.5 ± 1.7	87.0 ± 2.6
CDC 7027-3	76.9 ± 1.9	1.21 ± 0.05	1.41 ± 0.06	102.0 ± 5.7	81.2 ± 3.3	85.7 ± 2.1
CDC 7084-F1-7-62	77.4 ± 1.0	1.10 ± 0.03	1.25 ± 0.09	107.1 ± 4.0	86.1 ± 12.1	84.0 ± 0.0
CDC 7088-F1-21-1Y	81.5 ± 1.3	1.10 ± 0.05	1.00 ± 0.05	132.2 ± 12.6	84.8 ± 4.1	86.5 ± 0.5
CDC 6897-6	81.8 ± 1.8	1.14 ± 0.05	1.15 ± 0.01	68.7 ± 4.4	65.9 ± 5.5	79.1 ± 1.0
CDC 6948-4	81.6 ± 0.9	1.07 ± 0.05	1.19 ± 0.04	76.0 ± 3.5	75.7 ± 3.4	83.0 ± 1.7
CDC 7078-2	82.1 ± 1.1	1.15 ± 0.03	1.15 ± 0.04	84.4 ± 13.5	77.0 ± 5.7	84.7 ± 3.2
CDC 6999-2	80.8 ± 5.2	1.13 ± 0.06	1.25 ± 0.03	88.2 ± 10.0	81.7 ± 5.5	84.0 ± 1.7
CDC 7101-13	85.0 ± 1.3	1.15 ± 0.02	1.19 ± 0.03	77.1 ± 10.5	87.9 ± 3.8	86.3 ± 2.9
Mean	80.1 ± 3.7	1.20 ± 0.09	1.12 ± 0.12	95.5 ± 15.9	80.2 ± 4.8	84.4 ± 1.8
Range	70.5–86.8	1.07–1.43	0.95–1.41	68.7–132.2	65.9–87.9	79.1–87.0

between 20.8% to 27.5%. Recently, Galves et al. (2025) reported various pea breeding lines in Saskatchewan to have protein contents of 19.2%–28.2% depending on growing location.

### 3.2 | Functional Attributes of the Flours

The functional attributes of flour produced from 24 and 26 pea lines from the 2023 and 2024 growing season were reported in Table 2. Overall, solubility values were higher in 2024 (80.1%) compared to 2023 (76.2%). However significant variation among lines was evident, as values ranged from 68.7% (CDC Spectrum) to 86.3% (AAC Profit) in 2023; and 70.5% (CDC 6716-14) to 86.8% (AAC Profit) in 2024. CDC Spectrum had a lower solubility of 51.2% in Stone et al. (2021) than what was found for CDC Spectrum in the current study (68.7%–68.0%).

Water holding capacity was greater in the 2023 growing season (1.35 g/g) than 2024 (1.20 g/g). In 2023, CDC 6680-1 and AAC Lacombe represent the high (1.51 g/g) and low (1.26 g/g); whereas in 2024, the high and low were AAC Profit (1.43 g/g) and CDC 6948-4 (1.07 g/g). Our results are in-line with the WHC (1.2–1.4 g/g), reported for low-phytate and regular pea (CDC Bronco and CDC Amarillo) cultivars in Chigwedere et al. (2023), and CDC Meadow (1.13 g/g) in Stone et al. (2019). Galves et al. (2025) reported a mean WHC value of 1.7 g/g for both high and low protein lines (pea populations PR-25, PR-30, and PR-31) and a WHC of approximately 1.85 g/g for CDC Meadow. The WHC of CDC Spectrum in the current study was similar to that reported by Stone et al. (2021) (1.28 and 1.41 g/g vs. 1.37 g/g, respectively). Oil holding capacity (OHC) values were lower in 2023 (0.89 g/g) than in 2024 (1.12 g/g). In 2023, OHC values ranged between 0.86 g/g (CDC 6461-1) and 0.97 g/g (CDC 7088-F1-21-1Y); whereas in 2024, OHC values ranged between 0.95 g/g (CDC Forest) and 1.41 g/g (CDC 7027-3). Our values closely align with that of low-phytate and regular pea (CDC Bronco and CDC Amarillo) cultivars (OHC = 0.9–1.1 g/g) (Chigwedere et al. 2023). Stone et al. (2019, 2021) reported higher values (1.8–1.9 g/g) for CDC Spectrum, CDC Striker and

CDC Meadow. Galves et al. (2025) reported CDC Meadow to have an OHC of 1.3 g/g with similar results for both high and low protein lines.

Foam capacity (FC) and stability (FS) values were both higher in 2023 than in 2024. In 2023, the mean value of FC was 128.2%, with a high of 190.0% (CDC 6482-4) and low of 85.6% (CDC Spectrum); whereas in 2024 the mean FC was 95.5%, with a high of 132.2% (CDC 7088-F1-21-1Y) and low of 68.7% (CDC 6897-6). In contrast, Stone et al. (2019) reported FC of approximately 170% for pea (CDC Striker and CDC Meadow). Mean FS for 2023 was 88.2% and ranged between 84.7% (CDC 6308-3) and 92.6% (CDC 6844-9); whereas in 2024 the mean value was 80.2% and ranged between 65.9% (CDC 6897-6) and 87.9% (CDC 7101-13). Chigwedere et al. (2023) reported pea lines to have a lower FC (56.7%–68.9%) and FS (65.8%–77.5%) than the breeding lines in this study. All pea lines investigated in Galves et al. (2025), including CDC Meadow, had very high FC (200%–300%) whereas the FS was approximately 83%–92% depending on breeding line and location. The emulsion stability (ES) relates to how useful the ingredients will be in emulsified products including plant-based egg replacers and dressings or sauces. Overall, ES was similar regardless of the year with an average of 84.7%, with slight variation among lines. The breeding lines had higher ES (all > 80%) than what was reported for pea lines in Chigwedere et al. (2023) (all ≤ 70%), but lower than what was reported for CDC Striker (96.4%) and CDC Meadow (90.6%) in Stone et al. (2019).

### 3.3 | Training Machine Learning Models

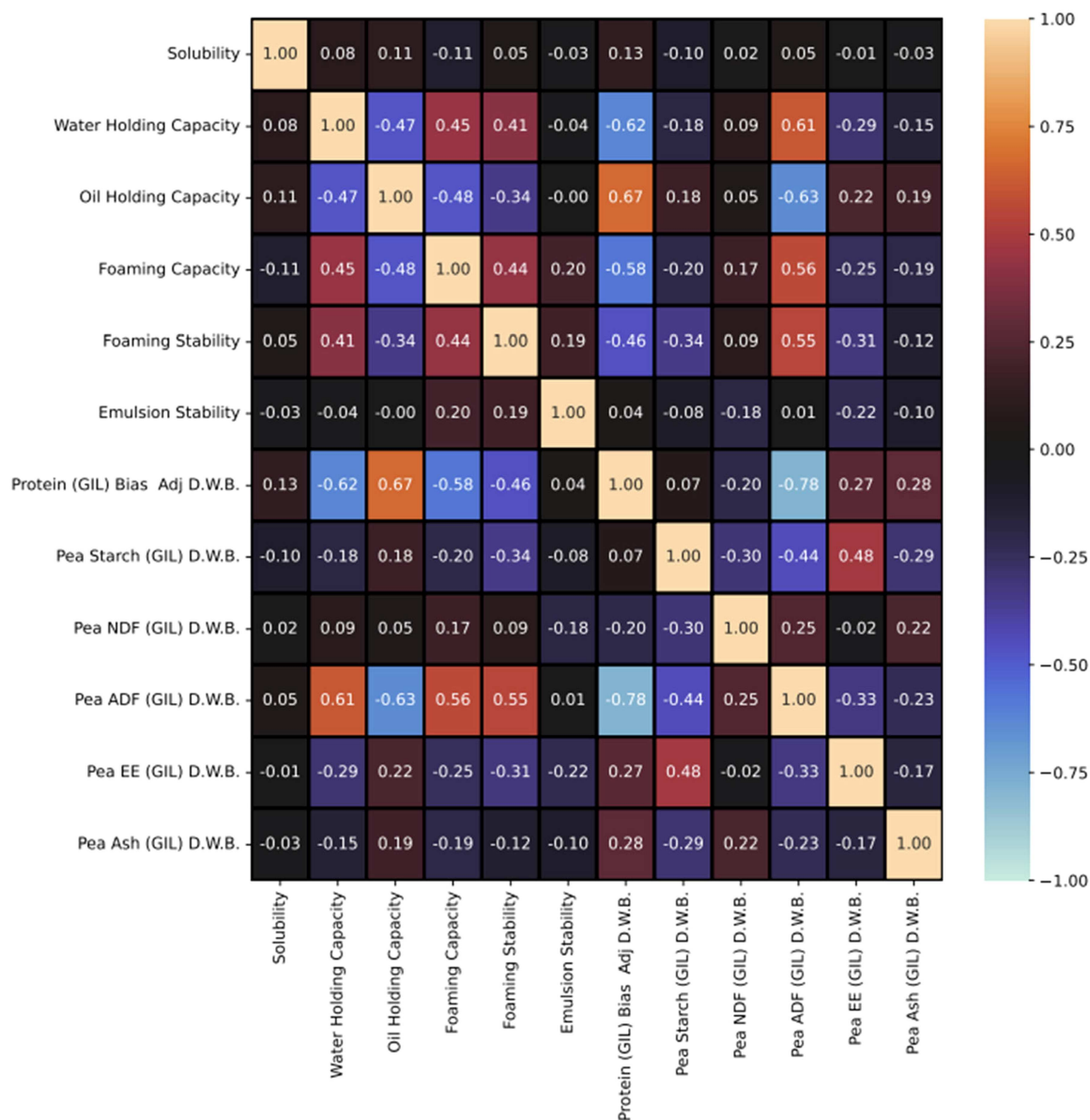
To aggregate the data for prediction, the mean was taken across replicates for each dependent and independent variable. Because not every breeding line was grown in both the 2023 and 2024 trials, the number of datapoints per measurement was either 3 or 6. Correlations between and among the sets of dependent and independent variables are shown in Figure 2. To evaluate the ability of each model to predict protein-related traits in unseen pea lines, we performed

leave-one-out cross-validation where one line was held out for testing and the rest were used for model training. This process was repeated for each pea line to obtain a final set of predictions. For each evaluation, outlier removal was performed on the training set by removing samples which were outside a multiple of 1.5 of the inter-quantile distance. Box-Cox power normalization was applied to each of the dependent and independent variables, with statistics for the transformation only calculated from the training samples. The exceptions were MLP, for which standardization was used instead for numerical stability, and XGBoost, for which no transformation was applied. Except for LR, each model's hyperparameters were tuned using leave-one-out cross-validation using the training set inside the outer cross-validation loop. The MLP model used a randomized search strategy with a budget of 100 iterations, while for the remaining models a grid search strategy was used. In order to avoid holding out data for early stopping, we trained the MLP model for 250 optimization steps, and controlled overfitting by searching over the learning rate instead. A summary of hyperparameters considered, and their search ranges are shown in

Supporting Information: Table S1. All models were trained using the mean squared error criteria.

### 3.4 | Validation of Machine Learning Models

For each scenario, the model's prediction accuracy was calculated in terms of mean absolute error (MAE), mean squared error (MSE), the coefficient of determination ( $R^2$ ), and Spearman's rank correlation (Spearman's rho). Prediction results are shown in Table 3. For additional interpretability results, individual conditional expectation plots, predictions, and residual plots are shown in Supporting Information (Figures S1–18). Across all evaluations, models consistently showed a moderate ability to predict water holding capacity, oil holding capacity, foaming capacity, and foaming stability, with  $R^2$  values between 0.398 and 0.45 for the best performing model. However, all models failed to significantly predict solubility or emulsion stability, with negative  $R^2$  values for each model.



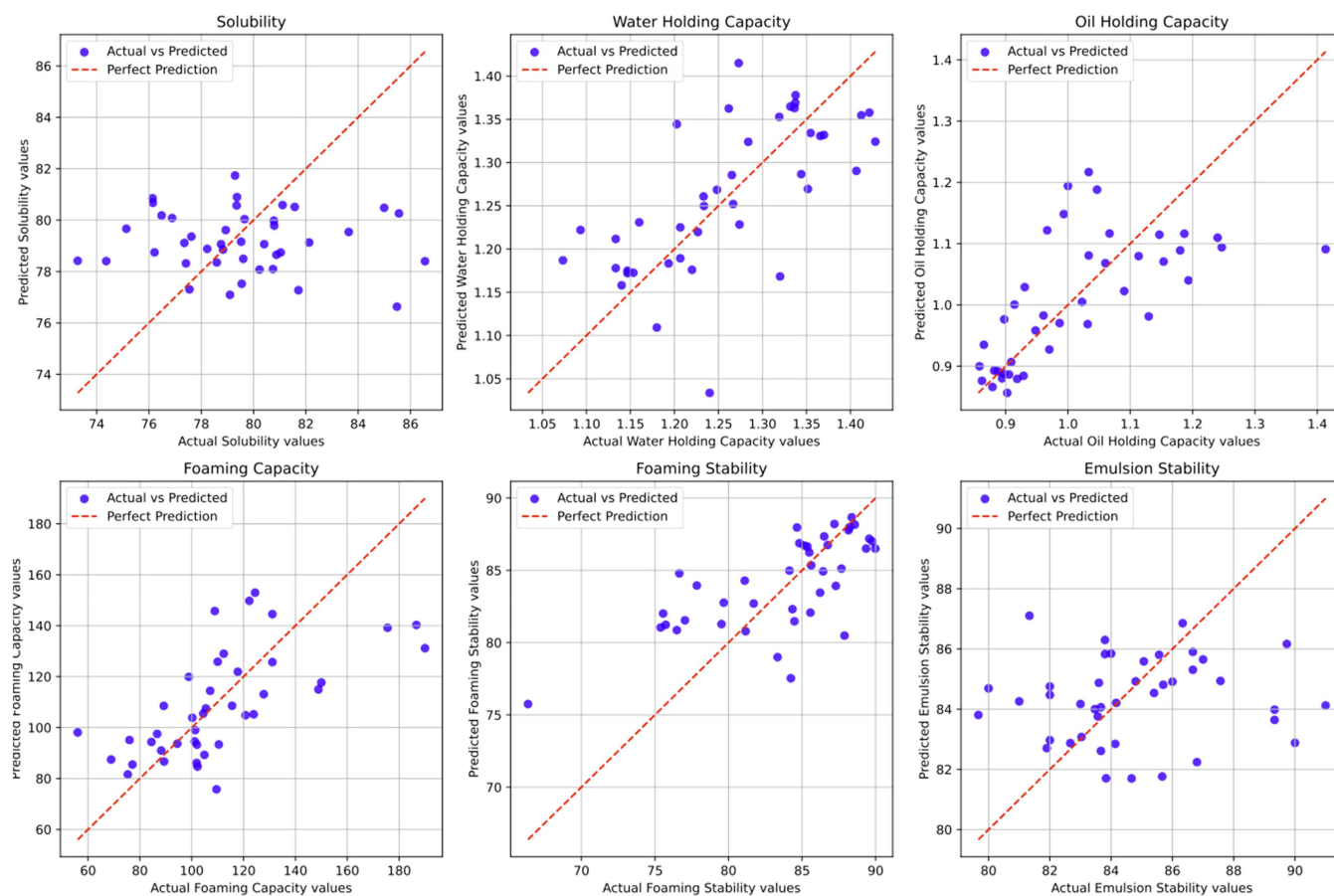
**FIGURE 2** | Pearson correlation analysis describing relationships between pea flour composition and ingredient functionality. ADF, acid detergent fiber; D.W.B, dry weight basis; EE, ether extract (i.e., crude lipid); NDF, neutral detergent fiber.

**TABLE 3** | Results for prediction of flour functionality.

Model	MAE	MSE	R <sup>2</sup>	Spearman's rho
(a) Solubility				
LR	2.517	10.990	-0.279	<b>0.078</b>
PLSR	2.497	10.470	-0.218	-0.046
SVR	2.288	9.129	-0.06	-0.006
GPR	<b>2.275</b>	<b>9.077</b>	-0.056	-0.896
XGBoost	3.109	15.454	-0.798	-0.106
MLP	2.474	11.004	-0.280	-0.391
(b) Water holding capacity				
LR	0.060	0.007	0.262	0.661
PLSR	<b>0.058</b>	0.006	0.373	<b>0.697</b>
SVR	0.063	0.006	0.262	0.647
GPR	0.061	0.006	0.318	0.641
XGBoost	0.058	<b>0.005</b>	0.431	0.647
MLP	0.060	0.005	0.329	0.664
(c) Oil holding capacity				
LR	0.074	0.011	0.352	<b>0.759</b>
PLSR	<b>0.073</b>	<b>0.010</b>	0.398	0.746
SVR	0.077	0.011	0.325	0.706
GPR	0.080	0.012	0.258	0.725
XGBoost	0.079	0.010	0.364	0.695
MLP	0.086	0.014	0.156	0.676
(d) Foam capacity				
LR	18.949	568.171	0.297	0.659
PLSR	<b>17.212</b>	<b>481.804</b>	0.403	<b>0.693</b>
SVR	21.744	997.811	-0.235	0.388
GPR	17.349	520.950	0.355	0.600
XGBoost	24.460	1028.412	-0.273	0.503
MLP	19.837	892.239	-0.105	0.418
(e) Foam stability				
LR	3.288	16.842	0.353	0.630
PLSR	<b>2.962</b>	<b>14.310</b>	0.450	<b>0.688</b>
SVR	2.983	19.626	0.246	0.634
GPR	3.550	24.219	0.070	0.506
XGBoost	3.861	25.869	0.006	0.550
MLP	3.551	19.818	0.239	0.574
(f) Emulsion stability				
LR	2.305	9.200	-0.296	0.103
PLSR	<b>2.194</b>	<b>8.634</b>	-0.217	<b>0.154</b>
SVR	2.264	9.342	-0.317	0.049
GPR	2.294	8.898	-0.254	-0.833
XGBoost	2.422	11.068	-0.560	-0.025
MLP	2.389	9.939	-0.401	-0.405

Note: The best value for each metric and trait is shown in bold.

Abbreviations: GPR, Gaussian process regression; LR, linear regression; MAE, mean absolute error; MLP, multi-layer perceptron; MSE, mean squared error; PLSR, partial least squares regression; R<sup>2</sup>, coefficient of determination; Spearman's rank correlation, Spearman's rho; SVR, support vector regression.



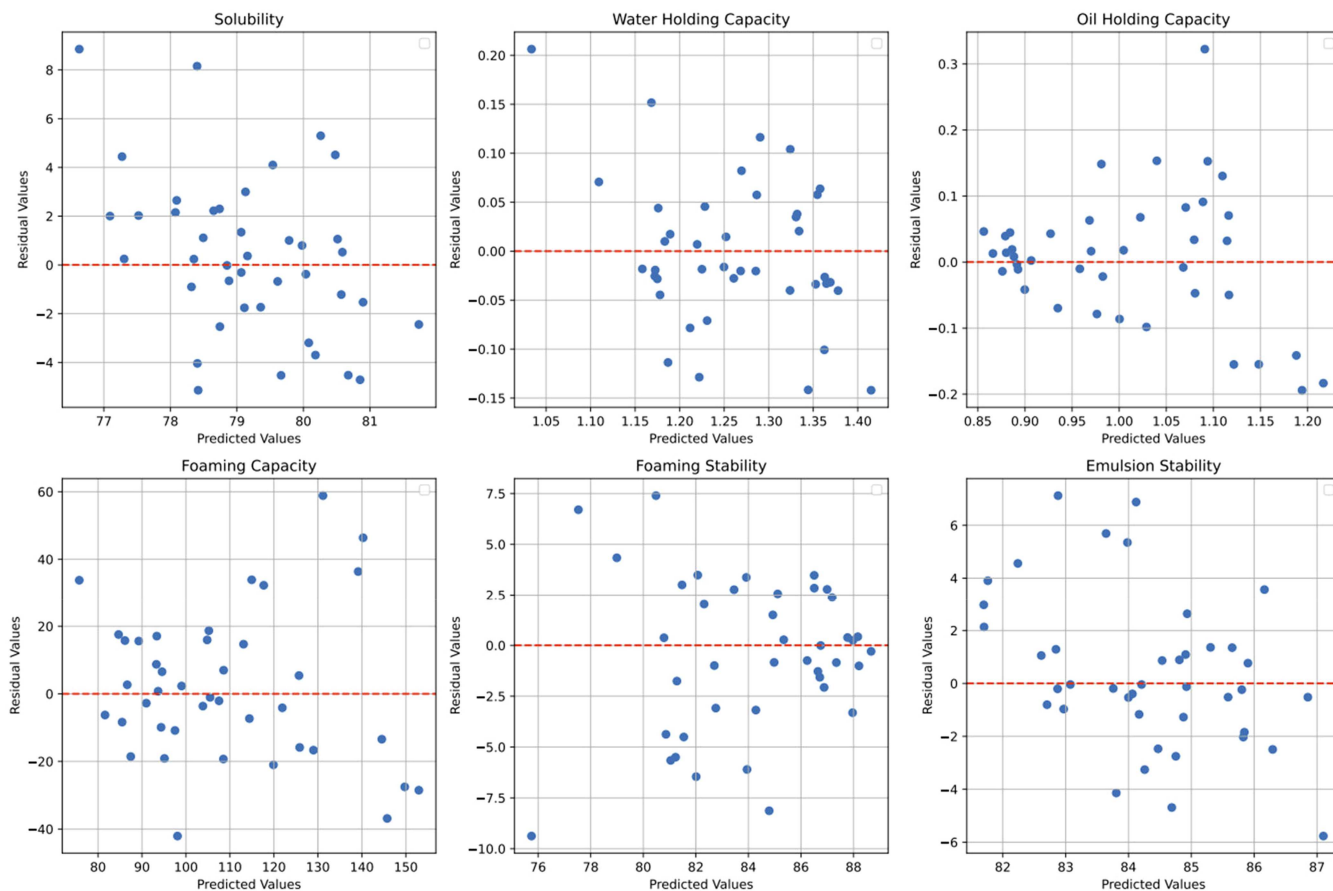
**FIGURE 3** | Partial least squares regression of predicted and actual values for each functional attribute of pea flour.

Among the models evaluated, PLSR demonstrated the best predictive ability across traits and performance metrics, with two notable exceptions (Figures 3 and 4). It underperformed GPR on solubility ( $R^2$  of  $-0.218$  vs.  $-0.056$ ), and it underperformed XGBoost on water holding capacity in terms of MSE (0.006 vs. 0.005) and  $R^2$  (0.373 vs. 0.431), while still leading in MAE (0.0581 vs. 0.0587) and Spearman's rho (0.697 vs. 0.647). PLSR also underperformed linear regression in predicting oil holding capacity, but only with respect to Spearman's rho (0.746 vs. 0.759).

In general, linear models tended to outperform non-linear models. For example, the MLP, being the model with the highest representational capacity, underperformed linear regression on five out of the six functionality traits. This is an expected effect of the small dataset size, as even with early stopping and high levels of regularization, high-capacity models such as neural networks often perform poorly on small datasets, due to their tendency to overfit the training data. The dominance of linear methods may also be because the relationships between composition and functionality are approximately linear in nature, and so these relationships are well-specified by linear models. The strength of PLSR over linear regression may be attributable to the fact that the composition covariates have a high degree of multicollinearity, which is better handled by PLSR. The poor average performance of XGBoost relative to linear regression on some traits ( $R^2$  of  $-0.273$  vs. 0.297 on foam capacity, 0.006 vs. 0.353 on foam stability) was more

surprising, as gradient boosted decision trees tend to perform relatively well on small tabular datasets.

To elucidate which composition traits contributed most to predicting functionality, Variable Importance in Projection (VIP) scores were calculated for each PLSR model (Figure 5). In general, scores above 1 are considered above average in their importance to the model's prediction, not just correlated variables. VIP scores consider variables within a multivariate model (i.e., considering all variables together), whereas correlations consider isolated one-variable relationships. This analysis shows that protein and acid detergent fiber (ADF) were both highly influential in predicting foaming capacity (1.52 and 1.55), foaming stability (1.30 and 1.54), oil holding capacity (1.64 and 1.50), and water holding capacity (1.56 and 1.53). These values complement Pearson correlation analysis which describes the relationship between protein and ADF values for predicting foaming capacity ( $-0.58$  and  $+0.55$ ), foaming stability ( $-0.55$  and  $+0.56$ ), oil holding capacity ( $+0.67$  and  $-0.63$ ), and water holding capacity ( $-0.62$  and  $+0.61$ ) (Figure 2). Overall, having more protein within the pea flour leads to poorer foaming properties, which is hypothesized to be due to a higher level of insoluble or aggregated protein, reducing its ability to migrate to the air-water interface to form a stable viscoelastic film around the air bubble. In contrast, higher levels of ADF would lead to increases in continuous phase viscosity which slows drainage to improve their foaming properties. Findings suggest that



**FIGURE 4** | Residual values from partial least squares regression for each functional attribute of pea flour.

both properties are of importance (driven by the high VIP scores), but foams are being stabilized more by increases in continuous phase viscosities from the ADF than protein-stabilized films at the bubble interface. In the case of oil holding, proteins have a positive impact due to the binding of oil to the hydrophobic amino acids on the protein's surface, which outweighs the negative impact of ADF which acts to dilute the protein matrix. In the case of water holding, ADF has a positive impact since it physically traps water within capillaries of the flour matrix and has good swelling behavior. Whereas higher levels of proteins, despite being hydrophilic in nature, are hypothesized to have a negative impact on water holding, since it dilutes the ADF levels and abides with less water than fiber does. Overall, both total protein and ADF are independent, their opposing (and competing) contributions drive pea flour functionality, and are both critical drivers within the PLSR models. To enhance foaming and water binding, the PLSR models suggest increasing the ADF content in the pea flours, whereas to enhance oil binding increasing total protein content is more important.

Protein bias was also highly important in predicting solubility (VIP score of 1.80), alongside starch (VIP score of 1.60). This was complemented by weak Pearson correlations between solubility and protein (+0.13) and starch (−0.10) (Figure 2). The multicollinearity findings suggest both protein and starch are indirect but crucial drivers in

predicting solubility, and most likely have a synergistic effect on solubility. This is hypothesized to be associated with the protein/starch ratio, where when the ratio is high, solubility would increase and vice versa.; however other components like ash, lipid and fiber could impact this complex mechanism. For emulsion stability, lipid content within the pea flour was highly predictive (VIP score of 2.02) with a weak correlation of −0.22 (Figure 2). The lack of a strong correlation alone suggests the total lipid content is an indirect driver of emulsion stability, where lipids might interact with other components (e.g., protein, starch, etc.) to impact stability such as the ratio of lipid to protein, starch and fiber, or combinations thereof.

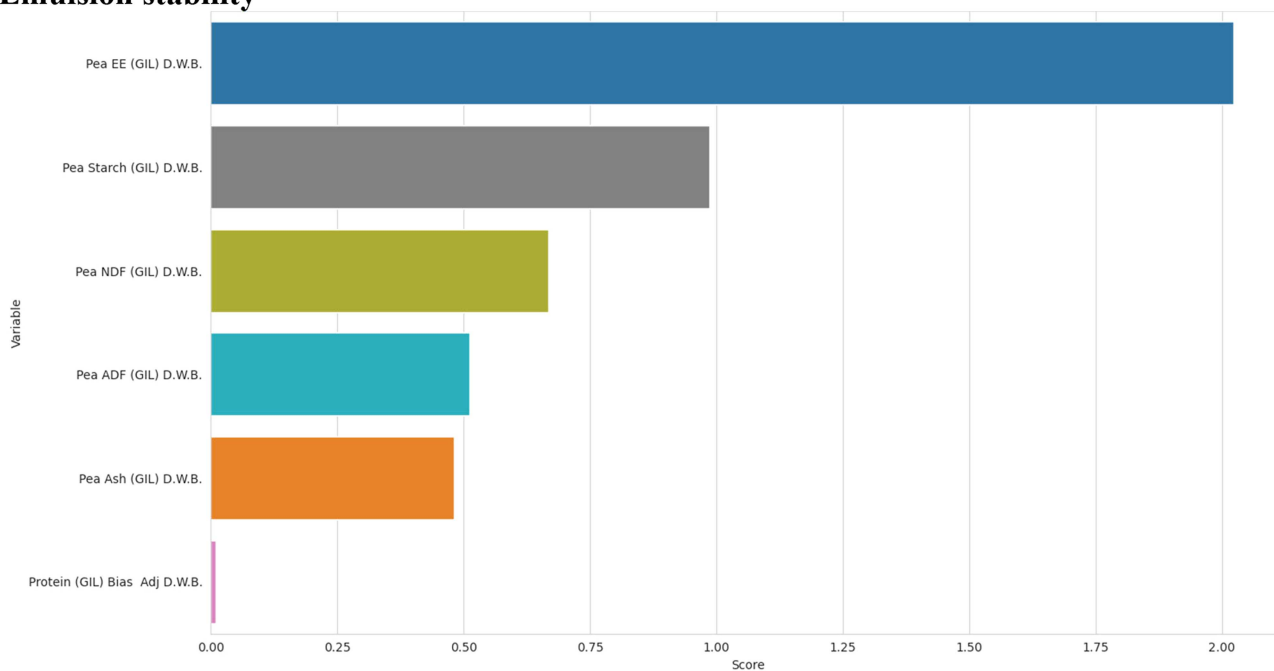
#### 4 | Conclusions

This study demonstrates that the predictive performance of machine learning models varies significantly across different functional properties of pea flour, reflecting the complex interplay among composition, genetic variation, and environmental factors. Among the diverse set of models evaluated, linear models performed better than the non-linear approaches, and PLSR offered the best predictive performance for most variables. For instance, the PLSR model predicted that protein and acid detergent fiber were both highly influential in predicting foaming capacity, foaming stability,

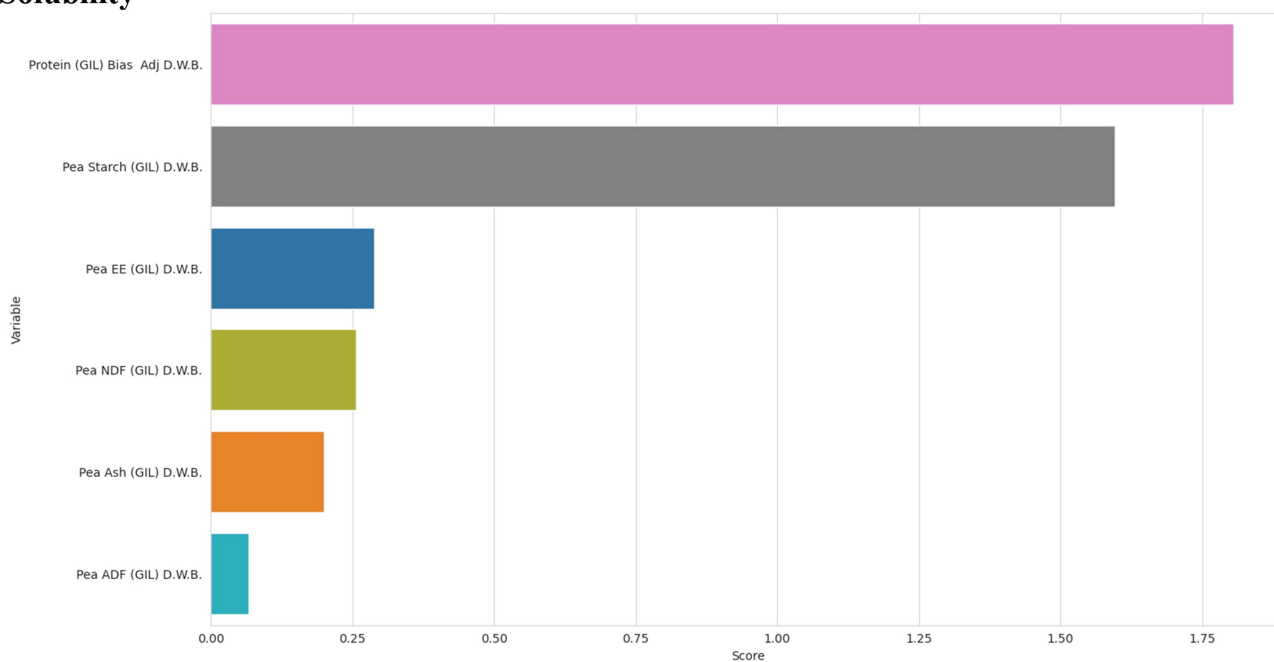
oil holding capacity, and water holding capacity as direct drivers (i.e., strong VIP scores with strong correlations). To have better foaming and water binding properties of the pea flours, the PLSR models suggest the composition of ADF should be higher (less protein), whereas for improved oil binding, flours should have higher protein (less ADF). However, the interplay between protein and ADF levels within the flour play a critical role in determining protein

functionality through competing mechanisms. In contrast, total protein and starch were also highly important indirect drivers (i.e., strong VIP scores with weak correlations) predicting solubility, whereas only total lipid was recognized as an important indirect driver predicting emulsion stability. For these functional attributes, the importance variables identified were influenced by complex relationships with other compositional parameters.

## A) Emulsion stability

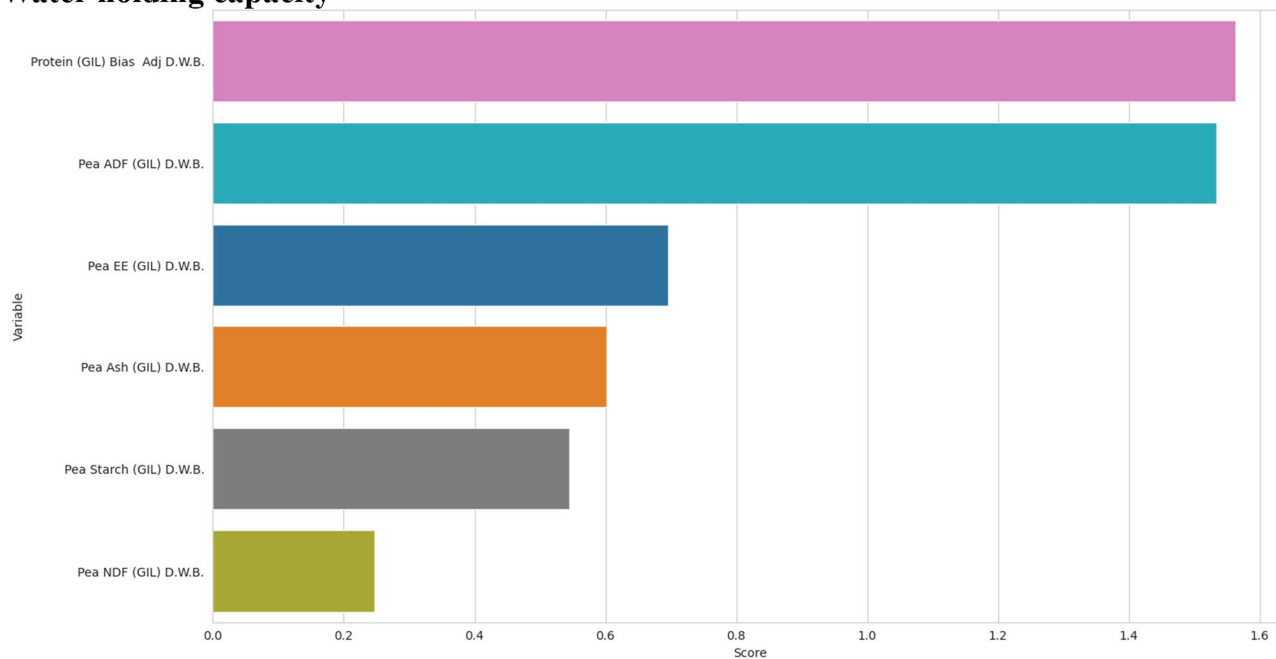


## B) Solubility

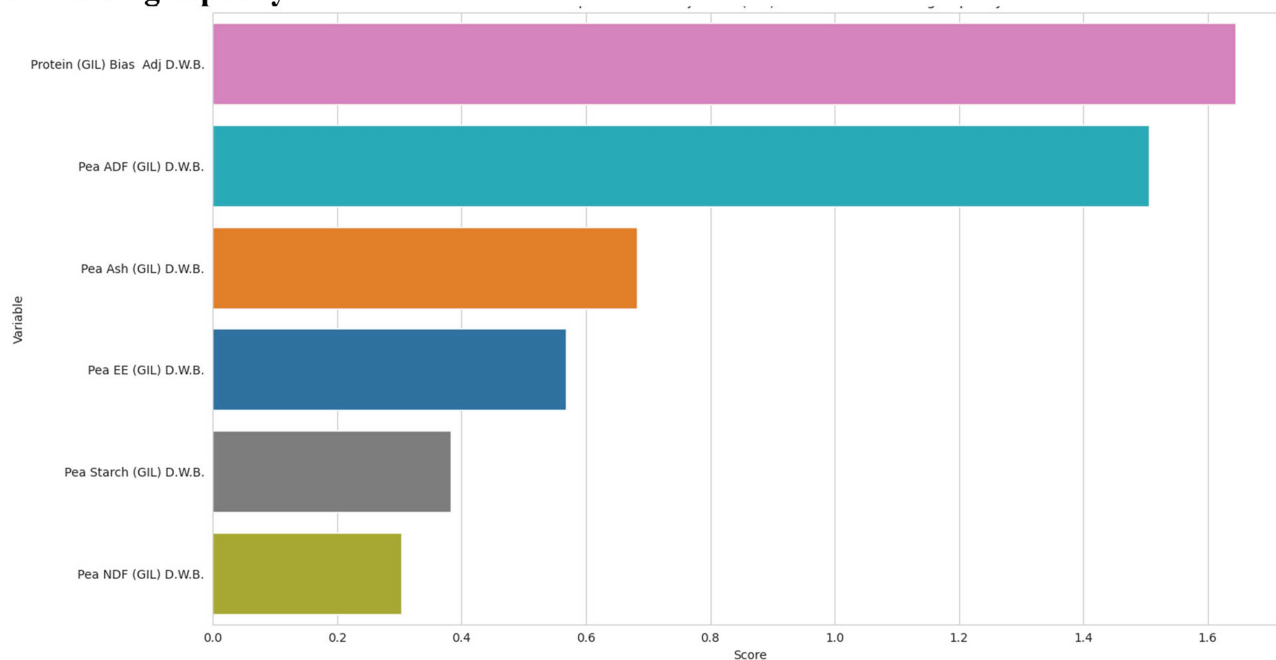


**FIGURE 5** | Variable importance in projection scores for (A) emulsion stability, (B) solubility, (C) water holding capacity, (D) oil holding capacity, (E) foam capacity and (F) foam stability of pea flours.

### C) Water holding capacity



### D) Oil holding capacity



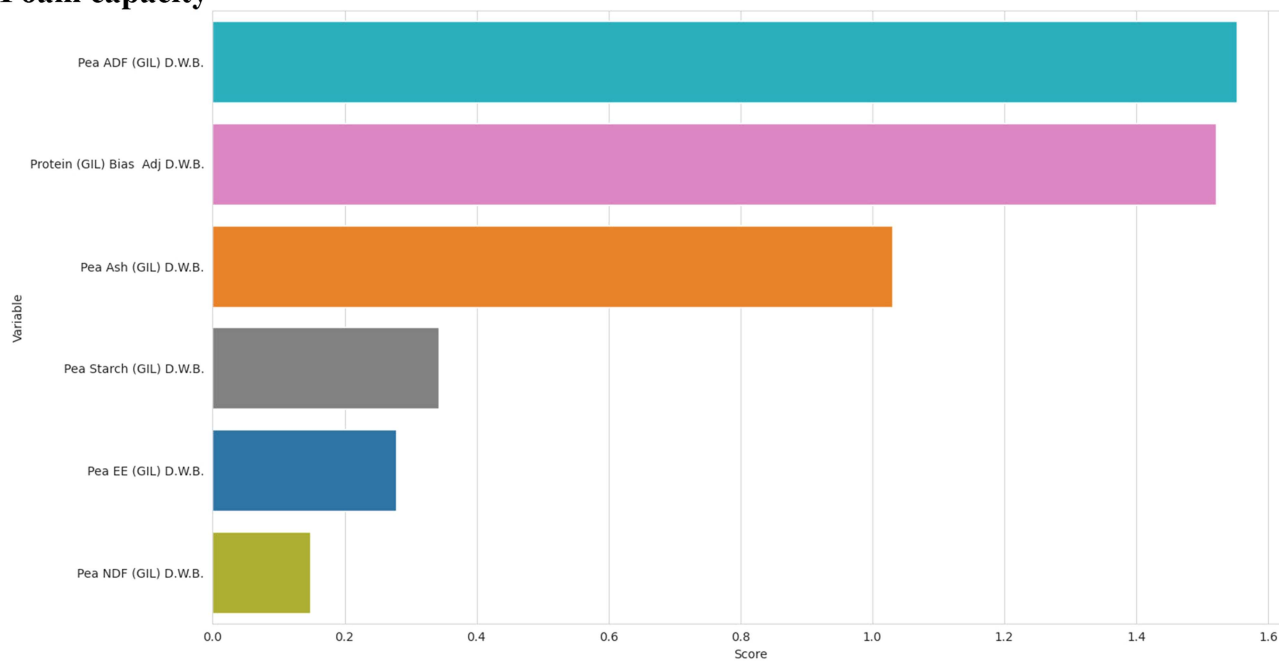
**FIGURE 5** | (Continued)

The ability to better predict the functional properties of pea flour (or other potential ingredients) based on composition will lead to more rapid ingredient selection for food product development purposes and help guide breeding programs for composition specific traits, reducing reliance on time-consuming and costly empirical testing. Furthermore, the capacity to link compositional traits with functional outcomes may provide breeders with actionable insights to guide the development of pea cultivars tailored for specific

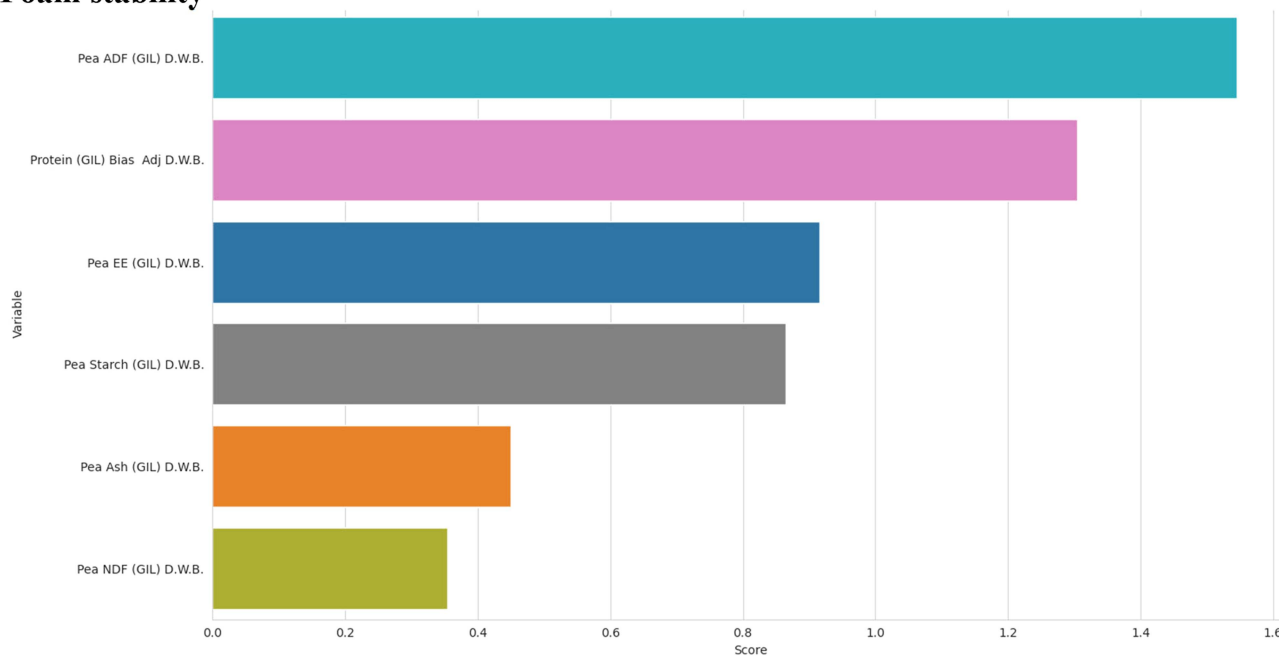
functional requirements, enhancing ingredient quality and consistency.

Overall, this work advances the integration of data-driven modeling approaches with food science and crop breeding, paving the way for more sustainable and efficient production of plant-based protein ingredients with optimized performance characteristics. Future research may focus on expanding datasets to include a wider variety of cultivars, environmental conditions, and processing variables, as well as exploring model interpretability and

## E) Foam capacity



## F) Foam stability



**FIGURE 5** | (Continued)

transferability across production batches to further enhance predictive robustness.

### Acknowledgments

Financial support for this research was provided by the Pulse Science Cluster under the Sustainable Canadian Agricultural Partnership and the Natural Science Engineering Council of Canada – CREATE Program. Technical support for machine learning was provided by National Research Council of Canada.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### References

Arganosa, G. C., T. D. Warkentin, V. J. Racz, S. Blade, C. Phillips, and H. Hsu. 2006. "Prediction of Crude Protein Content in Field Peas Using

- Near Infrared Reflectance Spectroscopy." *Canadian Journal of Plant Science* 86: 157–159.
- Chen, T., and C. Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1, 785–794. Association for Computing Machinery.
- Chigwedere, C. M., A. Stone, D. Konieczny, et al. 2023. "Examination of the Functional Properties, Protein Quality, and Iron Bioavailability of Low-Phytate Pea Protein Ingredients." *European Food Research and Technology* 249, no. 6: 1517–1529.
- Cordoba, H. A., R. Sadohara, K. K. Gali, et al. 2025. "Breeding for Plant-Based Proteins in Pulse and Legume Crops: Perspectives, Challenges and Opportunities." *Crop Science* 65, no. 4: 1–36.
- Dahl, J. F., M. Schlangen, A. Jan van der Goot, and M. Corredig. 2025. "Predicting Rheological Parameters of Food Biopolymer Mixtures Using Machine Learning." *Food Hydrocolloids* 160: 110786.
- Galves, C., K. K. Gali, T. Warkentin, J. House, and M. T. Nickerson. 2025. "High-And Low-Protein Pea Genotypes: A Comparative Study of the Composition, Quality, and Functionality of Flour." *European Food Research and Technology* 251: 1279–1288.
- Hara, P., M. Piekutowska, and G. Niedbała. 2022. "Prediction of Protein Content in Pea (*Pisum sativum* L.) Seeds Using Artificial Neural Networks." *Agriculture* 13, no. 1: 29.
- Hood-Niefer, S. D., T. D. Warkentin, R. N. Chibbar, A. Vandenberg, and R. T. Tyler. 2012. "Effect of Genotype and Environment on the Concentrations of Starch and Protein In, and the Physicochemical Properties of Starch From, Field Pea and Fababean." *Journal of the Science of Food and Agriculture* 92: 141–150.
- Kircali Ata, S., J. K. Shi, X. Yao, et al. 2023. "Predicting the Textural Properties of Plant-Based Meat Analogs With Machine Learning." *Foods* 12, no. 2: 344.
- Lie-Piang, A., A. Garre, T. Nissink, N. van Beek, A. van der Padt, and R. Boom. 2023. "Machine Learning to Quantify Techno-Functional Properties – A Case Study for Gel Stiffness With Pea Ingredients." *Innovative Food Science & Emerging Technologies* 83: 103242.
- Lo, B., S. Kasapis, and A. Farahnaky. 2022. "Effect of Low Frequency Ultrasound on the Functional Characteristics of Isolated Lupin Protein." *Food Hydrocolloids* 124: 107345.
- Nosworthy, M. G., S. Huang, A. Franczyk, G. C. Arganosa, T. D. Warkentin, and J. D. House. 2021. "Effect of Genotype, Year, and Location on the Proximate Composition and In Vitro Protein Quality of Select Pea Cultivars." *ACS Food Science & Technology* 1, no. 9: 1670–1676.
- Sharma, P., M. T. Nickerson, and D. R. Korber. 2024. "A Comparative Study of RSM and ANN Models for Predicting Spray Drying Conditions for Encapsulation of *Lactobacillus casei*." *Cereal Chemistry* 101: 1364–1379.
- Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks From Overfitting." *Journal of Machine Learning Research* 15, no. 1: 1929–1958.
- Stone, A. K., A. Karalash, R. T. Tyler, T. D. Warkentin, and M. T. Nickerson. 2015. "Functional Attributes of Pea Protein Isolates Prepared Using Different Extraction Methods and Cultivars." *Food Research International* 76, no. 1: 31–38.
- Stone, A. K., M. G. Nosworthy, C. Chiremba, J. D. House, and M. T. Nickerson. 2019. "A Comparative Study of the Functionality and Protein Quality of a Variety of Legume and Cereal Flours." *Cereal Chemistry* 96, no. 6: 1159–1169.
- Stone, A. K., S. Parolia, J. D. House, N. Wang, and M. T. Nickerson. 2021. "Effect of Roasting Pulse Seeds at Different Tempering Moisture on the Flour Functional Properties and Nutritional Quality." *Food Research International* 147: 110489.
- Wu, Q., X. Zhang, F. Gao, and M. Wu. 2023. "Study on the Residence Time and Texture Prediction of Pea Protein Extrusion Based on Image Analysis." *Foods* 12: 4408.
- Xie, C., M. Qiao, L. Yang, et al. 2024. "Establishment of a General Prediction Model for Protein Content in Various Varieties and Colors of Peas Using Visible-Near-Infrared Spectroscopy." *Journal of Food Composition and Analysis* 127: 105965.
- Zhu, L., P. Spachos, E. Pensini, and K. N. Plataniotis. 2021. "Deep Learning and Machine Vision for Food Processing: A Survey." *Current Research in Food Science* 4: 233–249.

### Supporting Information

Additional supporting information can be found online in the Supporting Information section.

**Supporting File 1:** cche70072-sup-0001-Supporting\_Information.docx.