Functionality-driven food formulation

Reducing the environmental impact using machine learning

Anouk Lie-Piang

Propositions

1. The 'mildness' of fractionation processes lacks an unambiguous definition.
   (this thesis)
2. Machine learning helps us capture the complexity of food.
   (this thesis)
3. Applied sciences raise the questions that shape fundamental scientific principles.
4. A manuscript ages like wine; it becomes better with time but only up till a certain moment.
5. Living in comfortable bubbles creates blind spots for an unsustainable lifestyle.
6. The celebration of small achievements is key to cope with perfectionism.

Propositions belonging to the thesis, entitled

Functionality-driven food formulation
Reducing the environmental impact using machine learning

Anouk Lie-Piang

Wageningen, 12 May 2023

Functionality-driven food formulation

Reducing the environmental impact using machine learning

Anouk Lie-Piang

Thesis committee

Promotors

Prof. Dr Albert van der Padt
Special Professor at the Laboratory of Food Process Engineering
Wageningen University & Research

Prof. Dr Remko Marcel Boom
Professor at the Laboratory of Food Process Engineering
Wageningen University & Research

Other members

Prof. Dr Vincenzo Fogliano, Wageningen University & Research
Prof. Dr Marianne Thomsen, University of Copenhagen, Denmark
Prof. Dr Frits Claassen, Wageningen University & Research
Dr Michel Vander Stappen, Unilever, Wageningen

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food and Nutrition Sciences),
(Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Functionality-driven food formulation
Reducing the environmental impact using machine learning

Anouk Lie-Piang

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 12 May 2023
at 1.30 p.m. in the Omnia Auditorium.

Anouk Lie-Piang

Functionality-driven food formulation
Reducing the environmental impact using machine learning
239 pages
PhD thesis, Wageningen University , Wageningen, the Netherlands (2023)
With references, with summaries in English

ISBN: 978-94-6447-584-5
DOI: https://doi.org/10.18174/587076

Content

Chapter 1

Introduction

Formulated food products are essential elements in many diets. Nonetheless, resources to produce them are becoming scarce. Therefore, we should reconsider the way we process foods ingredients and formulate food products. Food processing has been essential since pre-historic times as it ensures food safety and improves digestibility and overall quality of the food products.[1] A few of the most common and traditional processing techniques that aim to preserve foods are (sun-)drying, smoking, and salting. Fermentation is also a process to preserve fresh produce like milk into cheese or meat into sausages. Soaking and cooking pulses are necessary to remove anti-nutritional factors and increase nutrient availability.

In the 19-20th century, new technologies such as canning, pasteurisation, extrusion, and spray drying introduced another way of food preparation. To ensure food safety and quality, especially after the Second World War, home cooking shifted to more industrialized processes.[1] Further globalisation and urbanisation resulted in a larger demand for food products with less direct access to land. To make use of the economy of scale, industrial producers optimized food product formulations and increased efficiencies in the processes. One of the consequences was that refined ingredients are now preferred for processed foods like ready-to-eat soups and sauces since they have known and constant functional properties.[2]

Legumes, seeds, and grains are examples of common sources for such ingredients for food products. A traditional example is the usage of wheat flour for bread. In this case, the crop is processed to make it suitable for the final product: milling the flour breaks up the kernels and makes the components available for water to start the fermentation process. Currently, to obtain the standardized quality the crops are processed into multi-purpose ingredients, or isolates. This means that they are pure in one component. For example, a starch isolate is used to thicken a soup, a protein isolate is used to stabilize a plant-based yoghurt, and a fibre isolate is used to enrich pasta.

Although these ingredients were meant to increase the efficiency of the food chain from food manufacturers to consumers worldwide, they did not always result in efficiency over the full chain from crop to consumer. Refining crops such as wheat or legumes into almost pure ingredients and subsequently combining them again in

food products is not efficient (Figure 1.1). From a sustainability perspective, these products often contain ingredients that come from all over the world and have each already been processed extensively before assembling the final food product, which contributes to a higher environmental impact. For example, a study on meat replacers based on peas found that the ingredient preparation had the highest impact in the whole production chain.[3] In addition, fibre and other micronutrients are stripped from the ingredients, while the general intake of fibre is already low in our diets.[4,5]

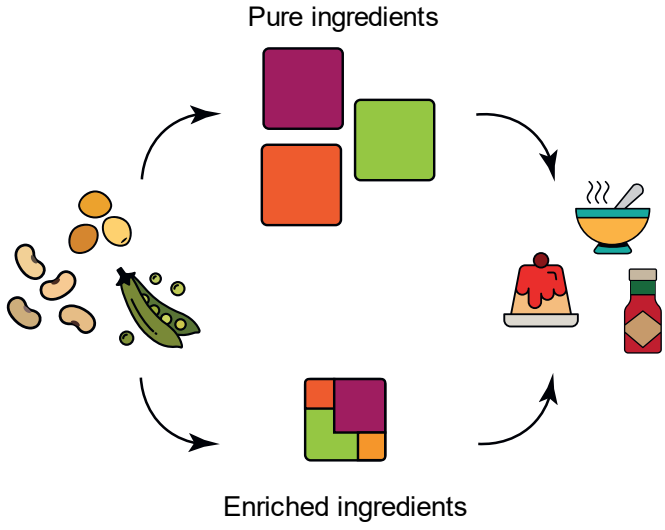

Figure 1.1 Fractionation of crops into pure and enriched ingredients (such as flour) to formulate food products. Colours of the boxes represent components in the ingredients.

The question arises if it is necessary to assemble food products using these highly refined ingredients. For some products, it is evident that purer ingredients are needed. For example, for vegan products or sports drinks that require protein to enrich the product. However, a simple bread or even a ready-to-eat soup can be made with less refined ingredients that already contain the combination of the components that were initially assembled by pure ingredients. Food assembly is therefore a trade-off between price, nutrition, functionality, and sustainability. In addition, food assembly cannot be generalized for all food products.

1.1 Moving to milder processes to produce ingredients

The production of highly refined pure ingredients has a relatively large environmental impact. For example, to isolate proteins, one typically disperses milled seeds (flour) into water and subsequently uses alkaline and acidic conditions to extract protein.[2] Although these methods do remove anti-nutritional factors[6,7], the consequence of this wet process is that the ingredients need to be dehydrated again for transportation, which is very energy intensive.[8,9] Besides, not all protein can be recovered, and a fraction is therefore lost. Finally, these harsh processing conditions change the functional properties of the proteins[10], which might be undesired. To step away from a focus on pure components, recent studies looked at creating ingredients that deliver certain functional properties, regardless of their purity.[2] With a focus on functionality, processes to fractionate crops into ingredients can become less intensive. For example, there are methods that use fewer chemicals[11] or no chemicals at all, such as sieving[12], air classification[13], electrostatic separation[14], or just water[15]. The fractions, or ingredients, that are produced using these milder fractionation methods can be highly functional, yet are only enriched in certain components, and far from pure (Figure 1.1).

It is the word functional that is important for food ingredients; however, it has at the same time also a very broad definition. Food ingredients have bio-functionality, which means for example that they include bioactive compounds with claimed health benefits. Digestibility can also be a functional property that is important, especially for foods targeted to specific groups, such as the elderly, infants or athletes.[16] Other functional properties that are relevant for food ingredients are techno-functional properties, such as gelation in for example puddings, emulsifying properties in mayonnaise, foaming capabilities in barista milk, or the viscosity in soups and sauces. Product developers formulate foods using ingredients with certain techno-functional properties in such a way that the structure and taste are optimized.

Although more mildly refined ingredients may have excellent techno-functional properties, they are relatively difficult to understand due to their complex composition. The latter can be mainly attributed to interactions that take place between components within the ingredient itself. The processing history of the

ingredients, which is mainly the fractionation method that is used to produce them, also influences the final techno-functional properties. Both the complex composition and techno-functional properties currently make these mildly refined ingredients less attractive to food product developers, even though they are more resource use efficient.[3,8,17] The central hypothesis in this work is that if these milder refined ingredients can be used in formulations for products like ready-to-eat foods and plant-based or enriched products, their environmental impact can be drastically reduced.

1.2 From functionality-driven fractionation to formulation

To create food products with a lower environmental impact, it is desirable to use more mildly refined ingredients, in spite of their complex composition and functionality. Jonkman et al.[18] proposed a multi-criterion decision framework in which ingredients are matched to final food products without being isolated into pure ingredients. This approach was inspired by the formulation of chemical products and was based on the principles of process system engineering. Here, blends are formulated based on their chemical compositions. Applied to food ingredients, complex compositions could be matched to a food product portfolio. Minimization of energy consumption over the production chain selected more mildly refined ingredients, but since only a subset of the fractions from the raw materials could be used, more raw materials were necessary. It was proposed that the applicability of (more mildly refined) ingredients will be increased when they would be selected based on their functional properties rather than composition (Figure 1.2). This means that regardless of whether the ingredient's composition is pure or complex, its techno-functional properties can be the same.

This poses the challenge of how to describe and quantify the techno-functional properties. There are many parameters that simultaneously influence the functionality of an ingredient, such as the processing history and interactions between components. Although mechanistic models can describe physical processes, the enriched ingredients have such complexity that it is very difficult to fully describe the functionalities using these types of models. The lack of understanding of all the interactions between parameters in models is not limited to food formulations.

Taifouris et al. [19] summarize well how in the design of formulated consumer goods (from perfumes to laundry detergents to food products), understanding interactions in both the molecular as well as the ingredient design limits the development towards optimizing formulated products. The need for optimisation models that can be used to quantify these non-linear systems is also highlighted.

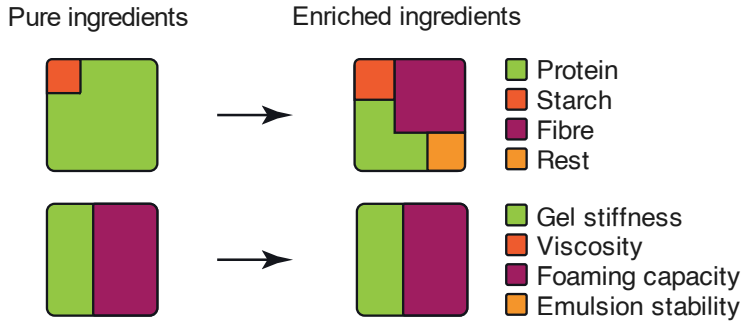

Figure 1.2 Schematic representation of ingredients in composition (top) and techno-functional properties (bottom).

Formulating food products based on functional properties using models has been done before. For example, Manzocco et al.[20] quantified the contribution of sugar replacements using explicit mathematical equations to formulate low-calorie syrups. A limitation of this study was that the interactions that took place between components were not covered. More complex mathematical models like machine learning algorithms are better capable of quantifying (non-)linear relationships between a large number of input and output variables, without prior knowledge of the physical behaviour. These machine learning models can be linked to mechanistic models to for example include the effect of the Maillard reaction.[21] Therefore, it is of interest to explore the quantification and prediction of techno-functional properties of mildly refined ingredients using machine learning algorithms.

1.3 Machine learning to capture techno-functional properties

In general, machine learning is a term used to describe mathematical algorithms that are based on empirical data.[22] Machine learning is widely applied in for example

speech and face recognition, language processing or finance for trading purposes.[23,24] It is also applied in the sciences, for example in social sciences to analyse political campaigns or in biology to predict protein folding and gene expressions.[25] In food science, it is increasingly applied in for example food waste prediction[26] and the classification of cow milk [27], which are just a handful of examples.

Machine learning covers a range of algorithms that quantify and predict relationships. Examples are linear regression, support vector machines, decision trees, and neural networks. Depending on the method used, the output can be a binary variable (e.g., spam or no spam), which is a classification problem, or a continuous data output (the optimal price to sell assets), which is called a regression problem. When in- and output variables are assigned to each observation, the types of models are called supervised learning.[28] An example is a study that compares a random forest, neural networks, and a support vector machine to classify different types of honeys based on the characteristics of an electrophoretic gel.[29] In unsupervised learning, model algorithms structure unlabelled data. For example, K-means clustering can be used to pre-process pictures of diseased leaves of corn plants, and to identify and remove noise for the picture and further analysis.[30] Another frequently used method is principal component analysis (PCA), which is used to reduce the dimensionality of high-dimensional datasets. This is useful to analyse multivariate data from for example sensory attributes of chickpeas stored under varying conditions.[31]

These examples demonstrate that machine learning covers a wide range of methods that can be applied to many fields. This can be attributed to the strong predictive capacity of these methods. As opposed to classical regression models, these models are flexible and can take any shape in which they are trained. This also poses the risk that the model overfits the data (Figure 1.3), meaning that the model follows every observation directly to obtain a good prediction. In other words, the models also fit the noise in the data. In that case, the model is too complex and it will become difficult to generalize or extrapolate the model to other observations.[32] When a model is simpler, it has the chance to underfit the observations. The causes of poor fits are complicated but can be assigned to noise in the datasets used to compose the

models or due to a suboptimal trade-off in model complexity.[32] Model complexity is dependent on for example the degree of a polynomial in linear regression, the number of nodes in a neural network or the number of splines used in a spline regression. These are the so-called hyperparameters of a model and set the architecture of the model.

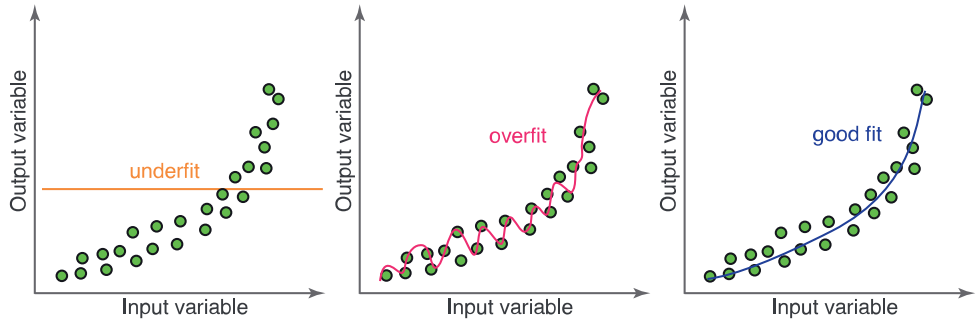

Figure 1.3 Representation of an underfitted, overfitted, and properly fitted model.

To determine the right model complexity and find the optimal architecture, the complete dataset is typically split into two datasets: a training set and a test set. On the training set, a k-fold repeated cross validation is applied. With this method, the training dataset is k times randomly split into a training and validation set and the error is determined for different splits (Figure 1.4). The error is calculated differently for regression and classification problems but in general, gives an indication of how well the data is predicted. For regression problems, this can for example be the square root of the sum of the squared differences between the measured and predicted values (root mean square error). For classification problems, this is often the percentage accurately predicted values of the total values.

The hyperparameters that determine the model complexity are then picked based on a trade-off between the bias and the variance of the model performance (Figure 1.5). The bias represents the ability of the training set to predict the training data, which is high when a model with low complexity (for example a low degree polynomial) underfits the training data and results in a high training error, and hence, a high

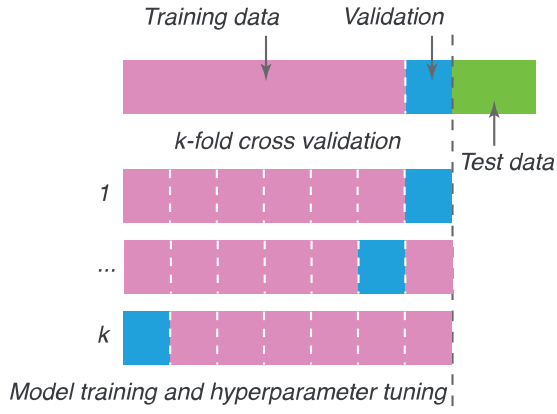

Figure 1.4 Representation of splitting the dataset into a training and test dataset, in which the training set is split k times into training and validation datasets to tune the hyperparameters, finally the tuned model is tested against the testing data.

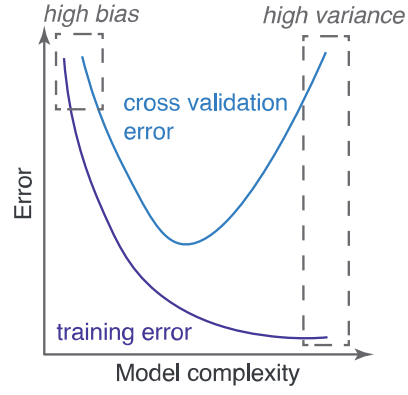

Figure 1.5 Bias variance trade-off of the errors of the training and cross validation sets as a function of model complexity.

training error, and hence, a high cross validation error. A model suffers from high variance when the model is very complex (for example with a high degree polynomial) and therefore overfits the training set, resulting in a low training error, but cannot generalize to the cross validation set that suffers from high error. The best model complexity is where both the training and the cross validation error are lowest. After the best model architecture is found, the model is subsequently tested again with independent data to evaluate the overall model performance. Other methods of building a model and tuning hyperparameters are leave-one-out, repeated cross validation, or bootstrapping.[28]

The large number of machine learning algorithms available with proven predictive capacity in many fields makes this a relevant field to connect to the formulation of food ingredients. Especially since understanding the underlying mechanistic processes is less important for this purpose since the aim is to identify parameter spaces that lead to certain functional properties in these ingredients.

1.4 Aim and outline of this thesis

This study is based on the hypothesis that the use of mildly refined ingredients may reduce the environmental impact of food products when selected based on their functional properties. To assess the complex composition and functionality of these ingredients, the high predictive capacity of machine learning algorithms can be used. Therefore, we connect these two topics in this thesis and employ machine learning in the formulation of ingredients for food products. The aim of this thesis is to assess to what extent the use of a functionality-driven selection of ingredients using machine learning may lead to more sustainable food production. This will be demonstrated using a case study of extensively and mildly fractionated ingredients. The ingredients will be produced from yellow peas and lupine seeds, as these are both excellent plant-protein sources. Ingredients from these crops will be matched to a product portfolio based on selected techno-functional properties: thickening behaviour (final viscosity and gelation), foaming capacity, and the stability of emulsions.

First, in **Chapter 2** the *state of the art* of fractionation methods is reviewed as well as the effect of extensive and milder processes on techno-functional properties of the ingredients. To quantify whether mild refining is actually more *sustainable*, the environmental impact of both extensively and mildly refined ingredients needs to be quantified. Therefore, in **Chapter 3** we conduct a life cycle assessment of extensively and mildly fractionated ingredients from yellow pea and lupine seeds.

The *concept* of a functionality-driven selection of ingredients is illustrated first in **Chapter 4**. Here, the thickening capacity of extensively and mildly fractionated ingredients is quantified and expressed in the final viscosity. The relation is subsequently used to create ingredient formulations that can be controlled for viscosity, composition, as well as global warming potential.

The functionality-driven selection of ingredients is then applied to the other selected techno-functional properties. As some of these properties behave non-linearly, other machine learning models that can quantify non-linear behaviour are employed. In **Chapter 5,** a *framework* is proposed to select machine learning models that deliver predictions that are physically feasible and that are suitable for ingredient formulations, illustrated by the gelling capacity of yellow pea ingredients.

As ingredient formulations commonly include ingredients from several crops, it is desired to have a general model that can be applied to multiple crops. In **Chapter 6**, all selected functional properties for *mixed crops* (both yellow pea, lupine seeds and mixtures of those) are quantified using the framework from Chapter 5. Based on all findings, a *general discussion* of the findings is provided with an example of the application of functionality-driven formulations of ingredients for a product portfolio in **Chapter 7.** This chapter gives a critical discussion of the sustainability of a product portfolio when selecting based on functional properties as opposed to composition reflects on the further implications of the presented concept and provides an outlook on further steps.

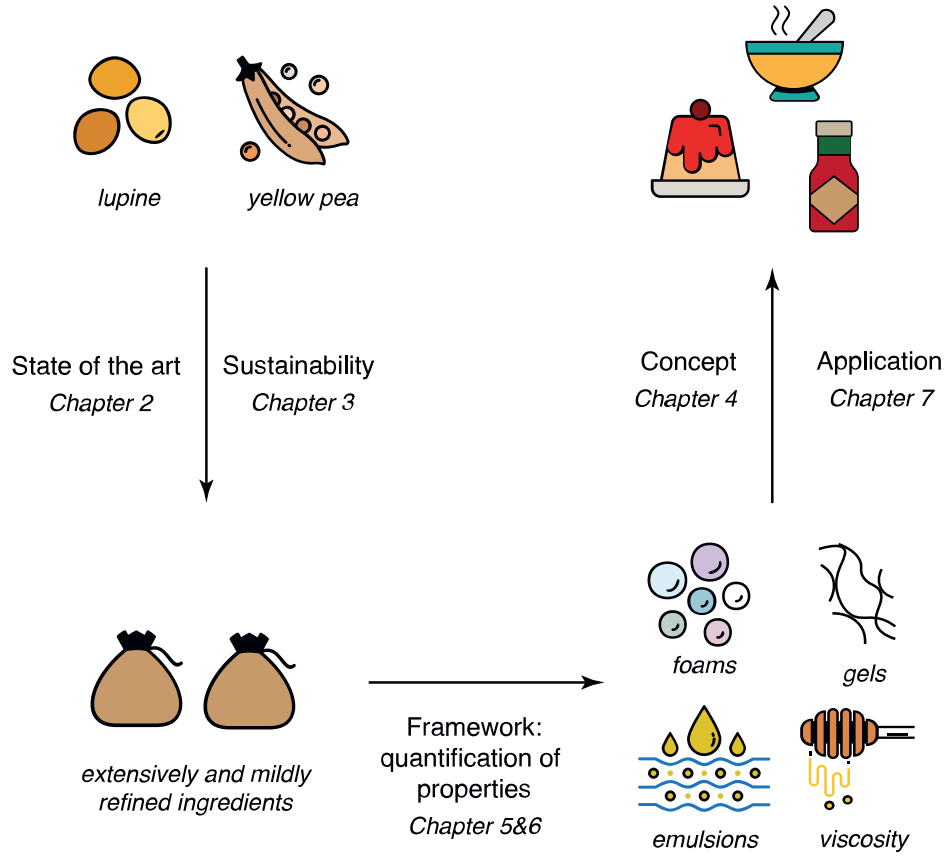

Figure 1.6 Graphical overview of the thesis content and chapters.

Chapter 2

State of the art

Mild fractionation for more sustainable food ingredients

A. Lie-Piang, J. Yang, M.A.I. Schutyser, C.V. Nikiforidis, and R.M. Boom

Annual Review of Food Science and Technology (2023)

2.1 Abstract

With the rising problems of food shortages, energy costs, and raw materials, the food industry must reduce its environmental impact. We present an overview of more resource-efficient processes to produce food ingredients, describing their environmental impact and the functional properties obtained. Extensive wet processing yields high purities but also has the highest environmental impact, mainly due to heating for protein precipitation and dehydration. Milder wet alternatives exclude, for example, low pH–driven separation and are based on salt precipitation or water only. Drying steps are omitted during dry fractionation using air classification or electrostatic separation. Benefits of milder methods are enhanced functional properties. Therefore, fractionation and formulation should be focused on the desired functionality instead of purity. Environmental impact is also strongly reduced by milder refining. Antinutritional factors and off-flavours remain challenges in more mildly produced ingredients. The benefits of less refining motivate the increasing trend toward mildly refined ingredients.

2.2 Introduction

Foods are generally created by combining ingredients into a complete food, which is greatly helped by the availability of ingredients that are neutral in taste and constant in quality. This requirement has led to the development of highly refined ingredients, such as oils, sugar, starch, and other carbohydrates like maltodextrins and protein isolates, which are relevant for the conversion into foods. Strong refinement of ingredients requires significant energy for heating and dehydration, and auxiliary chemicals such as salts, acids, and bases. It also inevitably results in the loss of some of the targeted components into waste or side streams. Processing therefore has a strong influence on the overall environmental impact, directly through the use of energy and chemicals, and indirectly through the requirement of more raw materials due to losses during processing.[3,9] Subsequently, there is a need to produce more raw materials.

The long-term perspective of food shortage comes from the growth and increasing affluence of the world population.[33] This is now aggravated by dramatically rising costs of energy and raw materials as well as unrest in some of the most productive agricultural regions of the world. More efficient production has therefore become paramount. At the same time, consumers tend to feel more negatively toward highly refined food components. Regardless of whether these worries are based on scientific evidence, it is crucial to take note of this. Finally, there is consensus among nutritional scientists that our diets do not contain sufficient dietary fibre.[4,5] The use of highly refined ingredients in our foods certainly does not help. The abovementioned trends all point to using milder processing methods, resulting in less pure ingredients that can be used for similar purposes as fully refined ingredients. However, mild processing may also have unwanted side effects, such as the retention of antinutritional factors[34] and microbial or chemical hazards in the ingredients.

Aside from proteins, which are currently in focus because of the trend toward plant-based meat analogues, another important type of refined ingredient is fats and oil. Because plant oils normally contain off-flavours and other unwanted components, extensive refining of oils has been the norm for decades. This process is typically

done by (cold and hot) expression of most of the oil and subsequent extraction of the residual oil using hexane. The oil is then degummed to remove phospholipids. The oil may be subjected to an alkaline step to remove free fatty acids and is then dewaxed to remove a high-melting oil fraction because consumers prefer oil that remains clear at all temperatures. The oil-refining steps lead to a loss of micronutrients.[35] Finally, bleaching by adsorption and deodorization by steam stripping yields oils that are neutral in taste but must be stabilized against oxidation.

Both protein isolation and oil refining are currently not efficient in the use of water, auxiliary chemicals, and energy, or in recovery. There are now sincere efforts underway to either create aqueous processes that employ milder conditions in terms of ionic strength and pH or avoid the use of water by using dry fractionation routes. These alternative routes all have different effects on the properties of the ingredients that are produced and compromise between the use of resources and the recovery that they can achieve. It is not yet clear how these alternative routes are related to the properties of the targeted fractions in terms of their technical functionality for incorporation into structured foods, nutritional quality, and retention of antinutritional factors and other hazards. Finally, a good perspective on the overall sustainability footprint of the different milder processing routes is also still missing.

In this review, we highlight the relation between milder processing routes for production of ingredients, resource use efficiency, and (techno-)functional properties of the obtained ingredients. First, different mild fractionation methods are discussed. We then focus on the functional properties that the various fractions have. Then, we assess the footprint of the different routes, and we conclude with a discussion of the trend toward milder fractionation processes and their potential for future applications.

2.3 Mild fractionation routes

Traditional wet separation often entails either the dissolution of part of the unwanted materials or dissolution and subsequent isoelectric precipitation of the desired component, such as protein, under elevated temperatures (Figure 2.1). Wet extraction enables the concurrent removal of antinutritional factors, such as

alkaloids[6] and trypsin inhibitors[7].However, the harsh process conditions are thought to disrupt the natural conformation and change the original properties of the proteins as present in the seed or bean. Therefore, a logical step is to consider using milder conditions, such as neutral or less alkaline pH, or a method that does not require water at all (e.g., dry fractionation).

Mild fractionation" for "food" is a relatively new term, which so far has been mentioned in only 14 published articles on Scopus, with the first published in 2014. Yet mild fractionation techniques have been used under different names. For example, dry fractionation has been applied for decades, for example, by Gueguen[36]. Therefore, it is no surprise that a search for "dry fractionation" for "food" yields 118 articles on Scopus, with the first published in 1984 and focusing on the separation of fats. From 2011 to the present, the focus has shifted to producing sustainable ingredients for consumers. Although these search terms give a good example of the novelty of the definition of mild or dry fractionation, they do not cover all the research going on in this field due to the inconsistent use of definitions. There are several excellent reviews on fractionation techniques themselves.[14,37–39] Here, we focus on the principles and the type of fractions that can be obtained using different types of fractionation methods that can all be considered mild.

2.3.1 Sieving and air classification

Fractions rich in protein, starch, or fibre can be obtained by dry fractionation using a combination of milling and separation of the flour particles in two or more fractions. Milling is the most crucial step, as it leads to a physical detachment between cellular microstructures that are rich in protein, starch, and other carbohydrates.[13] Ideally, individual starch granules, proteosomes, and bran fragments are created. Insufficient milling leads to suboptimal separation because the components have not been separated, whereas too strong milling leads to poor separation, as small particles aggregate through Van der Waals interactions. Crops with a high level of oil (>35–40%) exhibit liquid bridging by the oil and can be dry fractionated only after defatting. A recent study investigated the effect of oil content on pin milling of soybean and concluded that the exact milling procedure thus requires crop-specific optimization.[37]

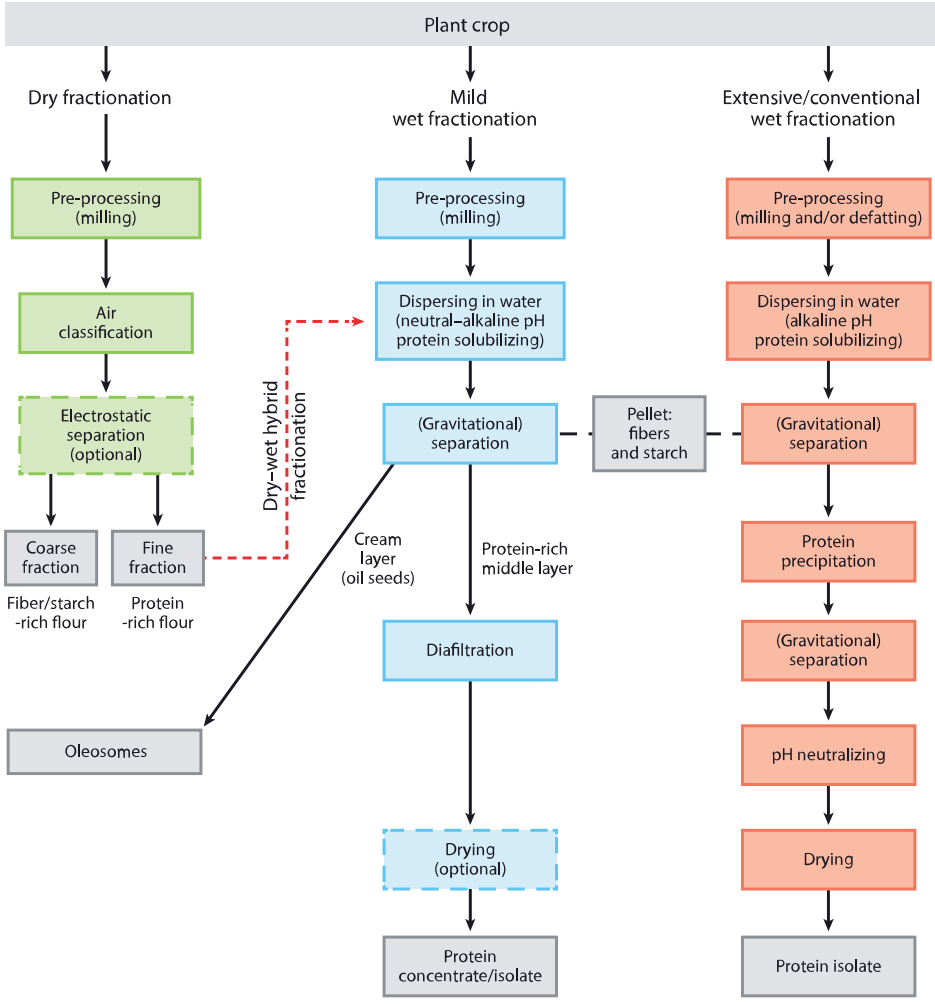

Figure 2.1 Schematic overview of dry fractionation, mild wet fractionation, and extensive or conventional wet extraction methods to produce plant-based ingredients such as protein- and starch-rich ingredients, oleosomes, or protein isolates.

Subsequent fractionation of milled flours can be done by sieving, elutriation, air classification, or a combination. Sieving can be done using a single sieve or stacks of sieves.[12] Elutriation is often operated in conjunction with sieving and employs an upward airflow that takes up smaller particles while larger particles settle downward.[38] Air classification with a classifier wheel is a further development of the elutriation principle and involves entrainment of the flour particles in an airflow, which is then blown through a classifier wheel. The advantage of all forms of dry fractionation is that all fractions are dry and therefore remain microbiologically stable. All fractions can be utilized for further application.

Air classification has been developing over many years for the enrichment of flours. Starch-containing legumes, such as field peas, navy beans[39], lentils, faba beans, cowpeas, and others[40], were found to be especially suitable for air classification. Yet (pseudo)cereals and oil-bearing crops like lupine seeds were also successfully enriched with air classification.[41–44] The reason that starch-containing legumes are rich in protein is explained by the cotyledon of these legumes, which comprises larger starch granules (±20 µm) that are embedded in a matrix of protein bodies (1–3 µm) and surrounded by a fibre-rich cell wall. These starch granules are liberated by impact or pin milling to enable size-based separation by air classification.

The emergence of crops as promising protein crops for human food and the high-energy-intensive production of ingredients has led to a re-emergence of the technique in the scientific literature from 2011 onward, for example, for barley[45] , oat[46]; corn[38], cottonseed, wheat, and soybean[47]; lupine[48]; and pea[49,50]. The process typically delivers fractions that are enriched in specific components, such as protein and starch, fibre, or other carbohydrates (e.g., arabinoxylans or glucans). Because the process does not require water, the energy consumption is an order of magnitude lower than that for wet processes.[51,52]

2.3.2 Electrostatic separation

Air classification separates components that have different combinations of particle size and density but cannot separate components that have similar size and density. For these flours, electrostatic separation may have potential, and the techniques have been described in previous articles.[14,53,54] The principle of the method entails

differential electrostatic charging of the powder particles followed by separation using an electrostatic field.

Electrostatic separation for food materials is relatively new. Plant materials that have been enriched by electrostatic separation are lupine[54], soybean[55], navy bean[56], rapeseed[57], oat bran[58], and wheat bran[59]. Ingredients produced from legumes were enriched in protein, whereas ingredients derived from oat and wheat bran were enriched in β-glucans and/or arabinoxylans. In other fields, such as the mining industry and waste recycling, triboelectric separation is the more established technology. For example, large-scale electrostatic separation is used to treat coal fly ash and mineral beneficiation, including beneficiation of iron, phosphate, talc, and calcium carbonates on a larger scale.[60]

Because the principle is different from air classification alone, good results can be obtained by carrying out air classification and electrostatic separation in sequence. Such 2D separations can result in protein concentrations of up to 65%, albeit after repeated electrostatic separation steps, as was shown in fractionation of yellow pea and lentil.[61] Xing et al.[61] enriched air-classified pea flour by electrostatic separation from 57.1% to 63.4–67.6% protein on a dry basis with an ingredient yield between 15.8% and 4.0%. More recently, researchers proposed the application of a magnetic field on top of the electric field to further enhance the separation of air-classified pea flour by electrostatic separation.[62] This could enrich air-classified pea from 59.7% to 72.1% protein content on a dry basis with an ingredient yield of 9.2%.

There seems to be a theoretical limit to the concentration at around 75% for yellow pea because this is the concentration of proteins in the proteosomes[63]; this limit is expected to be similar for other seeds. Concerning the relatively low yield of protein concentrate, it should be realized that generally electrostatic separation devices applied in reported studies are lab-scale devices not optimized for high yield. An advantage of dry fractionation is that all fractions (i.e., also starch- or fibre-enriched fractions) are not wasted as such but can be used depending on their functionality, such as thickening or water binding.

2.3.3 Mild wet fractionation

Boye et al.[64] provided a comprehensive overview of the main wet separations of plant proteins. Salt-induced or micellar extraction and precipitation avoids extreme pH values but employs a solution with high ionic strength to solubilize proteins. Subsequent dilution then precipitates the proteins. The remaining ionic strength supposedly protects the proteins from full denaturation.[10] Another strategy is to use just water for extraction. The yield of a water-only method is lower than with adjusted (more alkaline) pH or ionic strength, and thus the extraction is often repeated. The protein concentration varies from 54% for chickpea to 67% for pea, and therefore qualifies as a concentrate.[15] This process step was also applied to yellow pea to produce starch- and protein-rich fractions.[65,66] Ultrafiltration or diafiltration can also be used to concentrate proteins in the supernatant after mild wet fractionation. Möller et al.[67] applied multiple washing steps and ultrafiltration to the protein- and starch-rich fractions to increase the protein yield from approximately 78% to 87%. Another method without protein precipitation was applied by Kornet et al.[11,68] in which the proteins were first extracted using elevated pH, then purified and concentrated using ultrafiltration. They showed that avoiding a precipitation step by using ultrafiltration can recover the albumins from the raw materials, which normally do not precipitate because of their better water solubility and therefore are wasted. The fractions obtained by such a mild process yielded fundamentally different properties with respect to emulsion and foam stabilization[68] and gelation[11].

Besides protein and starch, milder processes to extract oil from crops are also in development. The isolation of oils from oil-bearing seeds is typically done by the expression and use of organic solvents, such as hexane. Hot pressing and solvent-based extraction make proteins less soluble[69] and oxidize other components in the seeds and therefore do not allow optimal use of the complete raw material. Because the use of organic solvents in the food process is under public scrutiny, the use of mild wet oil extraction is explored, as it allows the simultaneous extraction of oils, proteins, and other components.

Nikiforidis et al.[70] and later Romero-Guzmán et al.[71] explored the use of elevated pH and centrifugation to extract not just the proteins, but also oils in their native oleosome form. These are the intact vesicles that contain the oils in the plant cells and are still surrounded by a phospholipid monolayer, membrane-bound proteins, and generally some storage proteins.[72] Owing to their hydrophilic surface, they are dispersible in an aqueous environment. For instance, intact oleosomes were extracted by blending rapeseed and sunflower seeds in dispersion.[73,74] Ntone et al.[75] showed that both oleosomes and proteins could be extracted simultaneously, as well with elevated pH and centrifugation steps, and recovered separately. Moreover, Romero-Guzmán et al.[76] found that water with some ionic strength could extract these oleosomes, although the resulting emulsion flocculated depending on the cation used.

Regarding all extensive and mild wet processes, the thermal load by drying tends to denature or otherwise negatively influence the properties of the fractions. Proteins lose at least part of their solubility when dried[2], even in freeze drying[2,77]. Van Der Goot et al.[2] therefore implied that the omission of a drying step and storage and application as a liquid concentrate could benefit the properties of the materials as well as significantly reduce the overall environmental footprint of the extraction process.

2.3.4 Hybrid dry–wet fractionation

Even after air classification and/or electrostatic separation, the obtained fractions would qualify at most as concentrates rather than isolates. The use of a mild wet step specifically on the target fraction may help remove most of the residual fibre, starch, and other materials. This was shown for the separation of pea protein to be as simple as wet suspension and selective centrifugation of this fraction in water without pH adjustment.[52,66] The protein was concentrated in the top layers, which can be decanted and further processed by microfiltration and diafiltration to obtain a wet native protein concentrate of high purity. The hybrid procedure may yield purities in the range of those from isolates but at the cost of a somewhat lower yield. This hybrid separation has been evaluated for yellow pea, mung bean[78], and quinoa[79]. For quinoa, the wet phase separation was further improved by the addition of salt to

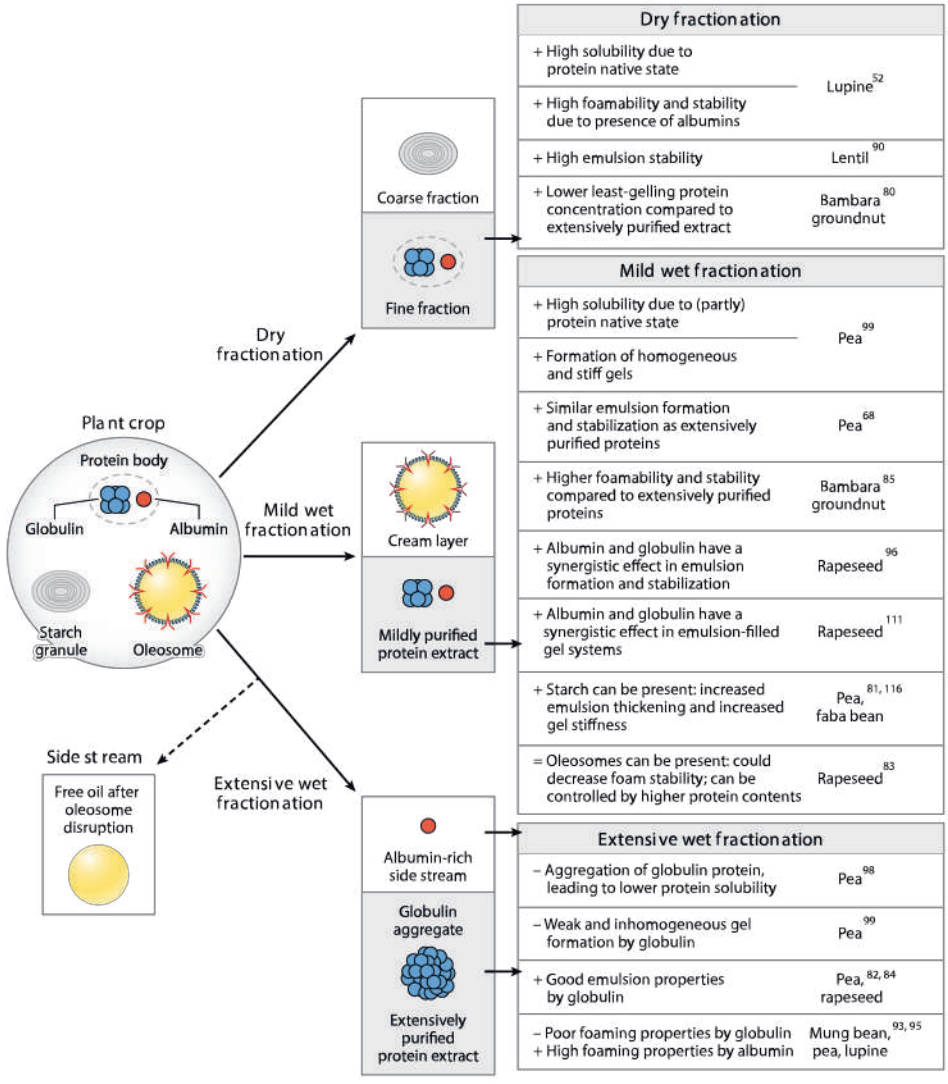

Figure 2.2 Overview of fractions that can be obtained with dry fractionation, mild wet fractionation, and extensive fractionation methods and their functional properties, such as solubility, foaming, or gelation.

enhance protein solubility and thus increase extraction yield. Interestingly, for faba bean, a hybrid fractionation process was proposed in which the coarse fraction after classification was subjected to a subsequent wet separation to increase protein recovery.[86]

2.4 Functionality of mildly fractionated ingredients

The (techno)-functional properties of mildly processed fractions are different from those of ingredients made with more intensive processing. In the following sections, the effect of different processing steps on protein, oil(-bodies), and starches and other carbohydrates is discussed. An overview of fractions obtained by extensive and mild fractionation methods is provided in Figure 2.2.

2.4.1 Proteins

The general focus in plant protein extraction is on the storage proteins[87], which are stored in spherical organelles called protein bodies or proteosomes[88]. Although wet methods disrupt these proteosomes, they can be extracted whole using dry fractionation methods.[49,89] Pelgrom et al.[48] showed that native air-classified lupine concentrates have better foam stabilization and similar digestibility compared to the heated protein concentrate. Dry-fractionated lentil protein extracts could be used to form stable oil droplets in emulsion systems.[90] Dry fractionation thus bears promise for obtaining protein extracts with good functional properties.

Protein purities around 90% can be achieved by wet extraction.[87] However, more extensive protein extraction generally leads to lower protein recovery.[91] In addition, the molecular properties (e.g., composition, structure, and solubility) can be affected by the extraction method. Generally, the focus in wet extraction is on water-soluble protein groups such as globulins and albumins. Globulins often exist as large quaternary structures, such as trimers and hexamers, whereas albumins are substantially smaller. In addition, globulins of many sources were reported to have a pI between 4 and 5, whereas albumins are soluble over a larger pH range (2–9).[92]

The conventional extensive protein extraction method is therefore mainly designed to extract plant globulins, as the alkaline pH solubilizes the globulins and albumins, but the precipitation step precipitates only the globulins. This leads to lower protein

recovery, as albumins may compose up to 25% of the total protein content of a plant seed.[93]

The separation of albumins and globulins influences the functionality of the ingredients, as they possess markedly different properties in, for instance, gelling, foaming, and emulsification. In the gelation of pea proteins, globulins can form stiffer gels compared to albumins, but albumins showed better gelation when mixed with dairy proteins.[94] Albumins were also superior for foaming, as they form very strong interfacial films (similar to dairy proteins), thus giving high foamability and foam stability.[93,95] Globulins only form weak interfacial films, leading to poor foaming. Albumins form more stable emulsions in a mixture with globulins than with single proteins.[96] Albumins are largely overlooked and should receive more attention in future studies. For example, the albumin-rich side stream has high value in (i.e., tofu or soy whey) tofu and soy protein production.[97]

The water-only protein extraction can recover both albumins and globulins at the same time. Dry fractionation of protein bodies also yields the coextraction of albumins and globulins, which explains the better foaming properties of dry-fractionated compared to wet-extracted lupine protein extracts.[52] Another advantage of mild extraction methods is higher protein solubility, as the precipitation step in wet fractionations generally leads to protein aggregation and thus lower protein solubility.[98,99] High solubility is essential for functionalities such as foaming and heat-induced gelation. However, heat-induced aggregation might be positive for extrusion. Therefore, the functionality of native proteins can be improved by heating, enzymatic hydrolysis, or chemical modification. For an extensive overview we refer to previously published work by Nikbakht Nasrabadi et al.[100]

Mild purification may lead to lower protein purities. (Dia)Filtration can remove other components to yield higher protein purity. Removal of phenols ensures high protein functionality, as they interact with proteins, leading to aggregation of the proteins.[101,102] Phenols impact taste, colour, and digestibility as well. Rivera del Rio et al.[16] investigated the impact of mild refining on the digestibility of pea proteins, measured with in vitro assays. Although alkaline process conditions reduce the digestibility, it was concluded that a heat treatment inactivates the trypsin inhibitors.

A mild thermal treatment increased the digestibility of initially insoluble proteins but decreased the digestibility of initially soluble proteins. The protein isolate was chosen as the benchmark and was not always the least digestible. However, an additional heating reduced the digestibility of the protein isolate even more, probably because of extensive aggregation of the denatured proteins.

In summary, the protein extraction method has a major impact on protein functionality but also on protein recovery. This relationship provides us with a unique tool, as the fractionation method can be used to tune the desired properties of the final protein extract.

2.4.2 Oils and oleosomes

Many plant seeds are mainly grown for their high oil content. The lipids are stored in storage organelles called oleosomes, also known as oil bodies or lipid droplets. Oleosomes are natural oil droplets of triacylglycerols that are surrounded by a monolayer of phospholipids with anchored membrane proteins.[103] In the current oil-extraction processes, the oleosomes are disrupted by pressing to extract the oil. The result is a pressed cake, from which oil can be further extracted using solvents (often hexane) to increase the oil yield. The remains are referred to as defatted meal (or cake), which is high in protein and often used for protein extraction.

Milder extraction of the oleosome extract could yield new possibilities in the formulation of oil-containing plant-based products while retaining protein nativity. The coextraction of oleosomes and proteins was shown for rapeseed[90] and Bambara groundnut[85]. Both materials were extracted at an alkaline pH (9–9.5), leading to highly negatively charged oleosomes and proteins, thus preventing interaction between the two components. A gravitational separation step resulted in three layers: (a) a top oleosome-rich cream layer, (b) a protein-rich middle layer, and (c) a pellet with insoluble material.

Several options are available here, as the supernatant (cream + middle layer) can be recovered as such, leading to an extract (milk) that is high in lipids and proteins. A second option is the separation of the cream layer and middle layer. The cream layer can be directly used as a dense emulsion system or diluted to desired concentrations.

Currently, oleosomes are disrupted to extract the plant oils, which are later homogenized with lecithin or plant proteins to create oil droplets in, for example, a cream. The disruption of oleosomes can be omitted as these are already naturally present oil droplets. Another promising feature is the oleosome membrane, which provides high stability against droplet coalescence, lipid oxidation, and heating.[104–106] The membrane proteins interact strongly with phospholipids by electrostatic and hydrophobic forces. Another functional property of oleosomes is the ability to take up free oil or encapsulate hydrophobic components.[107,108] Several studies addressed potential applications, such as oleosome-based emulsions[109] or pork-fat replacement with rapeseed oleosomes.[110] The protein-rich middle layer can be recovered and further purified by diafiltration to yield a protein extract with 65% protein and 15% oil.[75] A positive outcome here is a mild extraction of the proteins and the coextraction of both albumin and globulin proteins. This mild protein extract from oil seeds possesses good functional properties such as gelling, emulsifying, and foaming.[96,111,112]

2.4.3 Other components

Conventional extensive processing is generally oriented toward the recovery of a particular component, such as protein or oil. However, the conditions that give maximum yield of oil are not optimal for extraction of proteins: hot expression and solvent extraction degrade the proteins and may induce covalent bonds with polyphenols.[113] Avoidance of these conditions reduces the yield of the components of primary interest but increases the quality and yield of other components. This implies that mild processing is better suited for the total use of raw materials than conventional modes of processing.

Next to oil and protein, carbohydrates or fibre are the most important constituents. Karefyllakis et al.[74] stabilized oil droplets by using the fibre-rich fraction, which was obtained as an additional stream from (mild) sunflower protein extraction. Also, fibre-rich side streams could be utilized in bakery product applications.[114,115] Geerts et al.[116] used a starch-rich fraction to stabilize thickened oil-in-water emulsions due to the cooperation between starch and protein. A protein- and fibre-rich ingredient was produced from brewers' spent grain, which showed potential in a fibre-rich

pasta, with improved firmness and tensile strength. The presence of fibre also decreased the glycaemic index.[117]

With respect to all components, fractionation in general should not be aimed at isolating one particular component but should have the perspective of total use: fractionation of the raw material in as many high-value, functional fractions as possible. Although most attention has gone to the isolation of proteins and, to a lesser extent, oils, the refinement processes also yield other fractions, which can be valuable.

2.4.4 Antinutritional factors

The absence of antinutritional factors and off-flavours is important for food applications. Common antinutritional factors are protease inhibitors, amylase inhibitors, or lectins, whereas off-flavours, such as grassy or beany flavour, may be generated by enzymatic oxidation.[118] A potential disadvantage of milder fractionation methods is that the obtained ingredients are still raw and require further processing, e.g., heat treatment or fermentation, to remove antinutritional factors (ANFs) and tailor other properties such as taste and functionality. For instance, the beany taste of air-classified ingredient fractions is caused by lipoxygenase-catalysed oxidation, which can be prevented by a mild heat treatment.[52]

Heating or toasting is a widely used method to remove both off-flavours and antinutritional factors, but milder methods have also been developed. Dehulling of legumes can already remove some antinutritional factors, as reviewed by Asgar et al.[118] The type of milling and the settings of the classifier wheel speed can influence the concentration of antinutritional factors.[119] An example of a post-treatment on these ingredients is solid-state fermentation. Xing et al.[34] applied solid-state fermentation on air-classified chickpea flour using lactic acid bacteria. They could substantially reduce the antinutritional α-galactosides (raffinose, stachyose, and verbascose) by 88–99%, depending on the galactoside. The phytate level was reduced, and the resulting sourdough was stable and showed increased protein solubility. Fermentation is a sustainable and natural approach to removing

undesired components and can even generate compounds that positively influence the taste and/or nutritional value of the protein concentrate.

2.5 Sustainability assessment

2.5.1 Chain assessment

Both exergy and life cycle assessments (LCAs) are used to quantify sustainability of mildly refined ingredients. Exergy analysis bases resource efficiency on the first and second laws of thermodynamics[120], whereas a LCA focuses on the multidimensional impact of a process, ranging from the global warming potential to water usage[121]. Both consider the input of raw materials, energy, chemicals, and water. To quantify sustainability, the impact of each step in the production chain should be assessed (Figure 2.3). In general, plant-based ingredients have a similar supply chain, including cultivation of the crop, processing to final ingredient, and transportation. Next are food assembly and consumer usage, which are not discussed, as the scope of this review is on sustainable ingredient production.

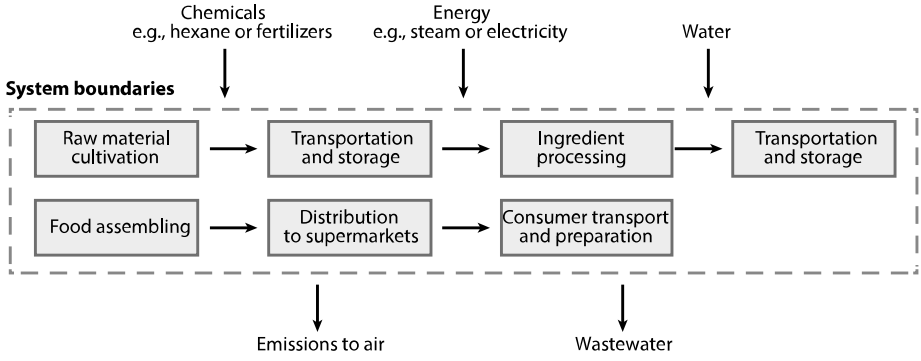

Figure 2.3 The production chain from food ingredients to consumer.

The impact of processing of the final product depends on the target component to be isolated. For example, the conventional wet extraction of starch and fibre from yellow pea[8] or potato[122] is relatively efficient, as it solely requires the suspension of the milled or ground crop with subsequent decanting, sieving, centrifugation, and hydrocyclones to separate the components. Drying of these fractions contributes significantly to the final resource use. The impact of drying is reflected in their share

of the cumulative exergy loss, i.e., destroyed useful energy, of 15% for the (pneumatic) drying of starches and 26% for the evaporator to concentrate fibres in the extensive wet fractionation method. This percentage is even larger in mild wet methods,[66,116] which emphasizes the importance of the impact of drying. Dry fractionation using air classification omits this, as drying is not required.[52] The impact of the omission of drying is projected in Figure 2.4, which shows that the extensive and mild wet production of a fibre- and starch-rich fraction have a higher impact compared to dry fractionation.[9]

On the other hand, wet extensive and mild isolations of protein have a higher global warming potential compared to the other components, due to protein extraction and

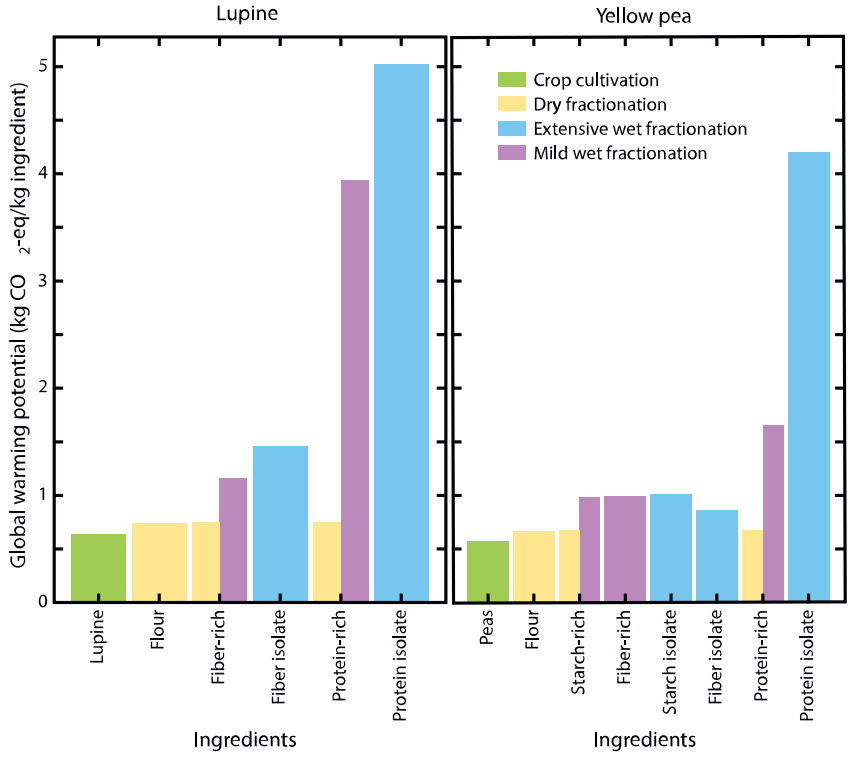

Figure 2.4 Global warming potential (system boundaries from cultivation to processing gate) of yellow pea and lupine of isolates produced with extensive wet fractionation and enriched ingredients produced with mild wet and dry fractionation. Data adapted with permission.[9]

drying (Figure 2.4). The latter contributes significantly to the overall impact and is responsible for 40% of the total cumulative exergy loss.[8] Because mainly the insoluble protein at the isoelectric point is extracted, there are many losses of soluble proteins. While isolates that are produced by isoelectric precipitation typically have a protein content of at least 79%, mild fractionation methods produce concentrates that are substantially lower in protein content. Mild wet fractionation yields approximately 54% protein, whereas dry fractionation yields 43% protein.[9] Nonetheless, the same study showed that the impact expressed per kilogram of protein instead of total ingredients is still lower for the more mildly refined ingredients than the extensively refined ingredients. More specifically, dry and mild aqueous fractionation of yellow pea yield a protein concentrate of 1.6 and 4.9 (combined for soluble and insoluble fraction) kg CO_2-eq/kg protein, respectively, and the extensive wet process yields a protein isolate of 5.3 kg CO_2-eq/kg protein. Vogelsang-O'Dwyer et al.[118] also calculated in an LCA that a faba bean protein isolate had a threefold higher global warming potential per kilogram protein compared to an air-classified protein concentrate.

When hybrid fractionation, a combination of dry and mild wet fractionation, is used to produce a fraction with just the soluble protein originating from the fine fraction, it has a protein delivery efficiency (ratio protein to invested life cycle energy) of 29.1 compared to 14.6 and 55.8 g protein per MJ for conventional extensive wet and dry processing, respectively.[52] The same difference was found for the soluble protein fraction from hybrid fractionation in a LCA.[9] Yet the side stream after centrifugation containing all insoluble components such as protein, fibre, and starch granules can also be considered as an ingredient. The same LCA study found that when considering the insoluble as well as the soluble fraction, the global warming potential per unit of protein comes closer again to the impact of a conventionally produced protein isolate. Nevertheless, hybrid fractionation is a relatively new concept and can still benefit from the optimization of water usage, as is done for mild wet fractionation separately.[67]

Lastly, the use of hexane and heat to distil the solvents during extraction of oil from crops leads to large chemical and physical exergy losses, which were shown in

lupine.[17] One could omit the oil extraction step, as it was found to have minimal impact on the functional property of lupine fractions.[123] The mild wet oleosome/protein extraction method discussed previously is another alternative. An exergetic comparison was performed on a final product containing oil, i.e., emulsions.[124] As the final application of these oleosomes is an emulsion, the exergy loss was compared between an oleosome- and conventional oil–based mayonnaise. The study showed that although the oil recovery efficiency is high in conventional extraction, there is a high chemical exergy loss due to the cake, which is only suitable for livestock feed. Mild extraction, however, requires a lot of water to process, but the efficiency, defined as the percentage of input exergy that remains useful after processing, of the mayonnaise based on oleosomes was 90% compared to 47% for conventional mayonnaise. Hetherington[125] calculated that an oleosome-based mayonnaise has the potential to be more environmentally friendly compared to conventional mayonnaise. Notably, both of these analyses are based on lab-scale experimentation; hence, yield or energy uses may be overestimated.[125]

Another essential element in the production chain of sustainable ingredients is the cultivation of crops and transportation to processing facilities. Comparing the impact of crops can be challenging because of the spread of production worldwide; hence, different distances exist between crop cultivation locations and the ingredient processing facilities. Nevertheless, the impact of cultivation can be compared to the impact of processing. For example, Heusala et al.[126] showed in an LCA that processing an oat protein concentrate contributes to a larger share of the total impact (75%) than cultivation (19%), attributable to the high energy costs. In contrast, for two faba bean concentrates, cultivation was responsible for a larger part of the total carbon footprint. Lie-Piang et al.[9] showed that cultivation is mostly responsible for the global warming potential of the production of dry-fractionated ingredients from yellow pea. However, for mild wet fractionation, the impact of processing for yellow pea and lupine is similar to and or larger than, respectively, cultivation. Both of these studies do not take the impact of food assembly into account, which can only mean that the impact of ingredient processing on the final consumer would be larger. This comparison shows that it is necessary to consider both cultivation and processing.

Earlier, we mentioned that drying is the largest contributor to all wet fractionation processes. However, it also reduces the environmental impact and costs of logistics and improves the shelf life. If the distance between the production and processing facilities is considered constant, the state of the product (wet versus dry) greatly influences environmental impact. Therefore, it is not desirable to transport wet products for long distances. However, for shorter distances under chilled conditions, transporting wet products might require fewer resources than drying the product first and transporting it in a dry state, depending on the final concentration of the product. For example, transportation of concentrated milk with 35% dry matter up to approximately 1,000 km still required less energy compared to the spray-dried variant.[127] However, there is a limit to the concentration of an ingredient for it to remain processable, which is probably relatively low (10–20 w/w%)[11] and thus requires a lot of energy for transportation and reduces its advantage. Another option to remove the drying step from fractionation methods is to produce ingredients in-house, just before their use in consumer foods. This latter option is made more realistic with the mild processing methods now developed, as they are relatively suitable for lower-scale production.

So far, the exergy and LCA studies presented in this review use either total ingredient or total protein to express the exergy efficiencies or impact categories. Even though both of these units are of importance, others might also be of interest. For example, the functionality that an ingredient will add to a product can also be expressed in environmental impact. This was presented in the functionality-based exergy assessment of Geerts et al.[8] who indicated a lower cumulative exergy consumption for the mildly fractionated ingredients per unit of functional property, in this case, viscosity compared to the conventional starch isolate.

2.5.2 Functionality-based fractionation and formulation

The final functional properties of ingredients depend on both the composition and type of processing. The latter relates to the structure that ingredients retain from their raw materials and also the degree of damage (e.g., thermal load or (de)hydration). One can adapt the processing routes to achieve the required functional properties instead of obtaining an ingredient of high purity. A previous

review by van der Goot et al.[2] introduced this topic and concluded that it is not always necessary to fractionate the crops into pure components and functional ingredients produced with mild fractionation have been used for decades, for example, with the use of flour in bread. In other words, the type and degree of intensity of a fractionation method can be adapted to what is required for the final application.

Möller et al.[67] showed that with additional washing steps in a mild aqueous fractionation, the purity of the protein fractions can be increased according to the final functionality in terms of protein content required. The isolation and extraction processes of yellow pea protein can be adapted to obtain fractions with excellent functional properties[11,99] while reducing the degree of processing. For oil-bearing crops, the extraction of oleosomes is an example of shifting the perspectives from extracting pure oil to attaining the oil bodies already as an emulsion.[109] Moreover, a more mildly derived protein mixture from sunflower seeds was also proven to efficiently stabilize oil/water emulsions and had similar functional properties to harsher refined sunflower protein isolate.[74] Another study found that a milder purification process of rapeseed proteins created more functional fractions, resulting in a protein extract with both napins and cruciferins, which are complementary in emulsion-filled gels.[111] The pH of mildly fractionated soy protein fractions can also be steered to obtain desired functional properties, which can result in, for example, a different water-holding capacity, protein solubility, or viscosity.[128] All these studies show that the fractionation methods can be adapted to achieve functional fractions and that the degree of fractionation should focus on the target functional properties required for the final food application.

Besides adapting fractionation processes of ingredients to attain specific functional properties, the focus in food formulation can also be on functional properties rather than purity. Jonkman et al.[18] proposed to create food formulations with a process systems engineering approach by blending different ingredients based on the target composition to increase the applicability of more complex and multicomponent ingredients. It was proposed that with a focus on functional properties, these blends could be made even more resource use efficient. Manzocco & Nicoli[20] formulated low-calorie syrups based on the functional contribution of each ingredient using

mathematical equations. In this way, food formulations could be optimized for caloric content while controlling functionality such as viscosity or colour in the final application. The mathematical equations were fitted to the functional properties of the individual ingredients, which was also the limitation of this study, as interactions between ingredients in formulations were not taken into account.

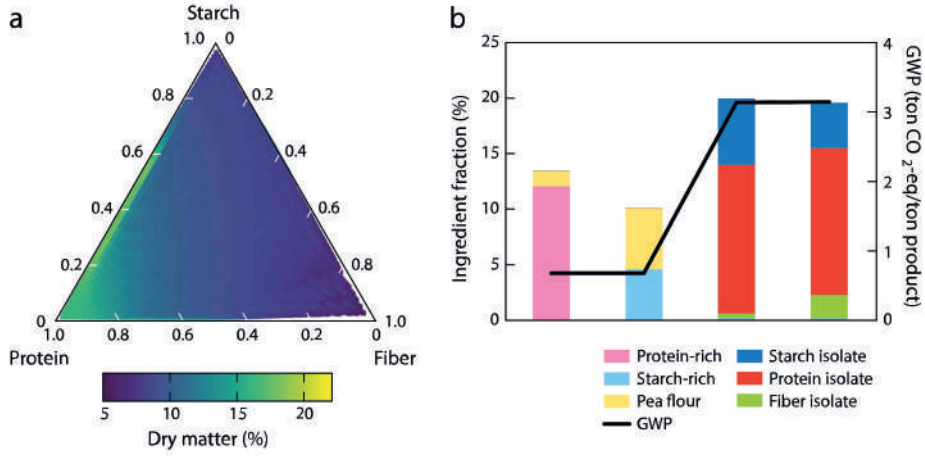

Figure 2.5 Formulation window of compositions that achieve 1,500 mPa.s viscosity at a shear rate of 160 rpm. (b) Possible ingredient formulations to achieve the same viscosity, including the respective global warming potential (GWP). Data adapted with permission from Lie-Piang et al.[129]

Lie-Piang et al.[129] extended this by creating ingredient blends, or formulations, based on functional properties. For example, the final viscosity of extensively and mildly refined ingredients was quantified using multiple linear regression. The advantage of using these types of mathematical models is that relations are quantified based on the data from ingredient blends, which means that possible interactions are covered. This allows the application of a broad range of ingredients with different functional properties and purities. For example, the regression model was used to generate the formulation window of components that achieved a similar viscosity of 1,500 mPa.s at a constant shear rate of 160 rpm (Figure 2.5a). The formulation window can be constrained to the composition of available ingredients with a different processing history and, hence, global warming potential (Figure 2.5b). The same viscosity could

be obtained with significantly less environmental impact when ingredients were used using mild fractionation compared with more extensively produced isolates. To apply this method to other (nonlinear) properties and consider process conditions as well, more complex machine learning algorithms can be used, such as neural networks.[130]

The combination of being able to adapt fractionation processes and ingredient formulations to achieve specific functionality allows us to minimize the environmental impact of the food products. At the same time, the functionality that is needed to create (healthy) foods is retained.

2.6 Conclusions and outlook

Although dry and other mild refining fractionation processes have been explored for many decades, they were originally considered to be cost-effective ways to enrich feed for livestock. The emergence in recent times of pulses, grains, and other plant materials as promising materials for plant-based foods for humans has led to a surge in the development of milder fractionation processes. The need for more sustainable ingredients is amplified because current methods for isolation are quite intensive in energy, water, and auxiliary chemicals and have relatively low yields.

There are two classes of mild processes, each with its advantages: dry processes and wet processes. The most important dry processes are air classification and electrostatic separation. The first separates based on particle size and density; the second—when based on tribo-electric charging—separates based on the surface properties of flour particles. They can be cascaded to enhance yield and recovery and obtain better separation between starch, bran, and protein. Dry separation at this moment is not suitable for crops that bear large amounts of oil (>35–40%) and thus first have to be defatted. The second class is mild wet separation, in which conditions are less extreme regarding the use of solvents, temperature, pH, and ionic strengths compared to the existing processes. The yields are generally somewhat lower compared to conventional alkaline extraction routes. Avoidance of a precipitation step leaves proteins in a more native state and improves their properties. In addition, mild wet separation can also help recover the albumins, which often do not

precipitate at the isoelectric point and are conventionally lost. These albumins have been found to have quite good technical functionality. The two classes can also be combined in hybrid processes, in which the first dry separation leads to an enriched fine fraction. Only this enriched fraction is then further purified, for example, by using a simple dissolution and sedimentation step in plain water. This leads to yields and purities of a protein fraction that are comparable to conventional wet isolation but with only a fraction of the energy and water used.

For oil-bearing crops, the mild wet routes seem promising, as they allow the recovery of oils as oleosomes, proteins, and other components. However, they still require significant amounts of water and either need dehydration, which is energy-intensive and compromises the quality of many fractions or result in aqueous concentrates that probably need to be chilled for storage and transportation.

As we are transitioning into an era in which raw materials, energy, and water will all be scarce, it is crucial to make the best possible use of the complete raw materials; we should therefore aim at using all different fractions. At the same time, environmental impacts of processes should be assessed to ensure that the most sustainable ingredients are used. This is especially important because the impact of processing in some cases exceeds the impact of crop cultivation. A focus on functional properties when designing fractionation processes or formulating ingredients will facilitate the use of more mildly refined ingredients.

Milder fractionation methods are very attractive for the intermediate future because they (a) can be cost-effective, as they are simpler than current methods; (b) deliver good quality ingredients; (c) are clearly more resource efficient and sustainable; and (d) fit the wishes of consumers for less processed foods. These milder processes are however still very much in development. The understanding and adaptation of technical and nutritional quality are only in their initial phases. In addition, the removal of antinutritional factors and other hazards are still an issue, even though several methods have already been developed for specific antinutritional factors.

Chapter 3

Sustainability

Less refined ingredients have lower environmental impact – A life cycle assessment of protein-rich ingredients from oil and starch bearing crops

---

A. Lie-Piang, N. Braconi, R.M. Boom, A. van der Padt

Journal of Cleaner Production (2021)

Abstract

In the coming decades, meat-based protein foods will increasingly be replaced by plant-based protein foods. These are typically prepared from highly refined protein isolates or concentrates, which require a lot of energy and auxiliary chemicals to be produced. Milder techniques such as dry fractionation or mild aqueous fractionation deliver alternative ingredients that have a multicomponent character but do not need the same amount of chemicals and energy. This study aims to assess the effect of reducing the degree of refining on the environmental impact using a life cycle assessment. The functional unit was 1000 kg of the processed crop. As protein is considered key in these ingredients, the functional unit of 1 kg of protein in the produced fractions was also assessed. The contribution of processing to the overall impact was found to be significant and, in some cases, larger than the contribution by the crop cultivation. Therefore, any analysis of the environmental impact should include both. Reducing the degree of refining substantially reduces global warming potential, human carcinogenic toxicity, fossil resource scarcity, and water consumption. However, for all impact categories, drying remains the largest contributor. The global warming potential of less refined ingredients was still lower compared to the conventionally refined ingredients when expressed per kg of protein, despite the significantly lower amount of protein. The fractions obtained through mild aqueous fractionation have a higher protein yield and a lower global warming potential compared to conventional full refining. Both dry fractionation and the combination of dry and mild aqueous fractionation substantially lower the environmental impact, but the protein yield and purity are also considerably lower. Overall, linking environmental impact to protein purity and yield allows for a comprehensive selection of sustainable food ingredients.

3.1 Introduction

Ready-to-eat soups, vegan ice-cream, and meat replacers are processed food products require texturisers that are often protein-based. These ingredients can be produced from animal products, such as whey protein isolate, or plants, such as pea protein isolate. Especially plant proteins are of increasing importance due to pressing climate issues, which require food producers to become more sustainable.[33] Currently, proteins that originate from oil-, starch-, or protein-bearing crops are isolated by isoelectric precipitation, requiring chemicals such as caustic (NaOH) and hydrochloric acid (HCL) for extraction and precipitation[131], and hexane in case of an oil-bearing seed.[132] Since every step in such a multi-step process involves intrinsic losses, the total yield of the desired components decreases with the number of steps. The side streams that are generated are often too dilute to use or may have such reduced quality that they are not suitable for human consumption anymore.[17] Therefore, we assert that to have adequate sustainability in our food production, it is important to utilize the complete crop and hence, the least number of resources.

Processing, or fractionation, methods have been developed that result in a lower degree of refining. These methods omit chemicals and aim at valorising the whole crop. Perhaps the most radical example is dry fractionation. The use of water is avoided, which normally dilutes the raw materials and inevitably takes some of the raw materials into the wastewater. Dry fractionation typically relies on differences in the size, density, and tribocharging properties of particles. This method includes milling and air classification or uses electrostatics to separate flours of pulses such as yellow pea, chickpea[13], or wheat bran[60] into protein- and starch-rich fractions. Another method is mild aqueous fractionation, which is a simplification of the conventional wet fractionation method – largely omitting the use of chemicals – and has been demonstrated for the oil-bearing lupine seeds[48] and soybean[128]. The same method was applied to the starch-bearing crop yellow pea.[65,98]

A lower degree of processing results in fractions that have a multicomponent character rather than being pure in one component. Nevertheless, promising functional properties have been found; such as the thickening capacity of mildly refined yellow pea[11] or soy fractions[128], and emulsifying properties of pea flour[133].

These milder fractionation methods for yellow pea[8] and lupine[17] have a better exergy efficiency compared to the conventional fractionation. Combining dry and mild aqueous fractionation can further reduce water and energy consumption. However, these exergy analyses only considered processing and did not include crop cultivation. Besides, they only considered exergy (useful energy) efficiency and therefore did not consider other modes of environmental impact. Lastly, these methods for lower levels of refinement create fractions with a lower protein content, while it is considered the most important component for which more sustainable alternatives must be found. To evaluate these three factors, we here report on a full attributional life cycle assessment comparing different degrees of refining; based on the whole process and protein content, from cultivation to processing.

A life cycle assessment (LCA) allows for a multidimensional assessment of the environmental impact of products at all stages in the food production chain. This includes the extraction of resources and emission of hazardous substances.[121] Most current LCA studies focus on calculating the footprint of the production of only conventional protein isolates[134] while less have focused on the milder fractionation methods[135]. Therefore, this study aims to directly compare the environmental impact of protein-rich fractions from conventional and milder fractionation methods from starch- and oil-bearing crops. It is hypothesized that a lower degree of refining decreases all sustainability indicators due to the lower use of resources. Further, expressing the environmental impact in protein will negatively influence the footprint of a lower degree of refining as their products contain less protein than the fully refined isolates. To exemplify the effect of processing on the impact of protein-rich ingredients, a starch- and an oil-bearing crop were selected: lupine seeds and yellow pea. These crops were selected due to the significant protein content of approximately 40[6] and 23%[131] respectively. In addition, these crops have established functional properties such as foaming abilities[136], emulsifying capabilities[98], and gelation[99]. At first, the most significant environmental impact categories of processing these crops will be presented. After this, the relative impacts of processing and cultivation will be compared. Finally, the impact of the individual process steps and the fractions relative to the purity and yield of protein is discussed. Evidently, the selection of sustainable ingredients is not merely based on the purity of the

fraction, as other factors such as functional properties are also important. However, these are not within the scope of this study.

3.2 Goal and scope

The goal of this study was to quantify the impact of decreasing the degree of refining on the environmental impact of food ingredients using an attributional LCA. To achieve this, four different refining processes were compared for yellow pea and lupine, namely: conventional protein extraction (CF), mild aqueous fractionation (MF), dry fractionation (DF), and combined dry and mild aqueous fractionation, or hybrid fractionation (HF).

3.2.1 Ingredient selection

In this study, the production of ingredients from lupine legume seeds (*Lupinus angustifolius L.*) and yellow pea (*Pisum sativum L.*) in the Netherlands was considered. The yields of the ingredients from the starting material and the purities were obtained or based on information of the processes in literature (further explained in 3.3.1) (Table 3.1). In the case of hybrid fractionation of yellow pea, the protein content was determined experimentally, for which the dry fractionation by air classification[51] was combined with mild aqueous fractionation.[98] The protein content of the ingredients was determined using Dumas analysis (Nitrogen analyser, FlashEA 1112 series, Thermo Scientific, Interscience, Breda, The Netherlands). A conversion factor of 5.52 was used for the calculation of the protein content.[137] Generally, the milder fractionation methods lead to ingredients that contain less protein but could have a higher protein yield. The latter is defined as the protein in the ingredients as a percentage of the original protein present in the crop and can be calculated using Equation (3-1).

$$\text{Protein yield (\%)} = \text{Ingredient yield (\%)} \cdot \frac{\text{Protein content of ingredient}}{\text{Protein content of crop flour}} \qquad (3\text{-}1)$$

Ingredient yield is the mass of the obtained ingredients as a percentage of the initial mass of the starting material and the protein content is the amount of protein in the specific ingredient or crop flour.

3.2.2 System boundaries

The system boundaries were set from cultivation to the end of ingredient processing, in other words: cradle-to-processing-gate (Figure 3.1). These boundaries were picked because the food assembly, distribution, use of the product, and disposal after the production were considered not relevant for the current comparison between high and low degrees of refining. The analyses in this study were performed both including and excluding the impact of the cultivation and transportation, to highlight the effect of decreasing the degree of refining in processing only. Energy, chemicals, and water that go into the system were considered for this analysis, as well as the emissions to air and wastewater that come out of the system, which is further explained in the life cycle inventory in section 3.3.

3.2.3 Functional unit & allocation

The functional unit was defined as 1000 kg of the processed crop to compare the impact among the different processes. This functional unit was picked since the goal of this study was to assess the overall impact of reducing the degree of refining on the complete process. Since protein is the key component of the ingredients, 1 kg of protein as a functional unit was also investigated. The allocation was done based on mass (dry matter) allocation. Fractions that were considered as a loss (i.e., soluble solids after precipitation), were treated as waste; hence no impact was allocated. This means that all useful co-streams produced during the development of the protein-rich ingredients (i.e., starch or fibre-rich fractions) received an allocation according to mass in all analyses. In this study, when discussing the functional unit per 1000 kg crop, the footprint of all fractions was considered. For the discussion of the environmental impact per kg protein, only the allocated impact of the protein-rich ingredients was considered.

3.3 Life cycle inventory

The data collected for the agricultural production and transportation of yellow pea and lupine were retrieved from the Agri-footprint 5.0 database.[138] For more details on how this data is derived one can consult the database description, which is publicly available. The impact for all individual fractionation processes was calculated using data for electricity, process steam, water, and chemicals retrieved

Table 3.1 Total yield mass ingredient from starting material and protein content in the ingredients based on dry matter (db%) of the protein-rich fractions obtained through conventional, mild aqueous, dry, and hybrid fractionation.

Crop	Method	Fraction	Ingredient yield (db%)	Protein content (db%)
Yellow Pea	Raw material	Yellow pea flour [66]		21.4
	Conventional fractionation	Protein isolate[66,131]	22.3	78.8
	Mild aqueous fractionation	Soluble protein[98]	23.9	55.9
		Insoluble protein[98]	10.2	53.3
	Dry fractionation	Fine fraction[66]	22.8	42.9
	Hybrid fractionation	Soluble protein	5.1[52]	62.2*
		Insoluble protein	3.8[52]	37.4*
Lupine	Raw material	Lupine flour[17]		39.5
	Conventional fractionation	Protein isolate[17]	27.0	87.0
	Mild aqueous fractionation	Enriched protein[17]	29.0	86.5
	Dry fractionation	Fine fraction[17]	33.0	57.6
	Hybrid fractionation	Enriched protein[17]	14.0	87.0

*Obtained experimentally by this study

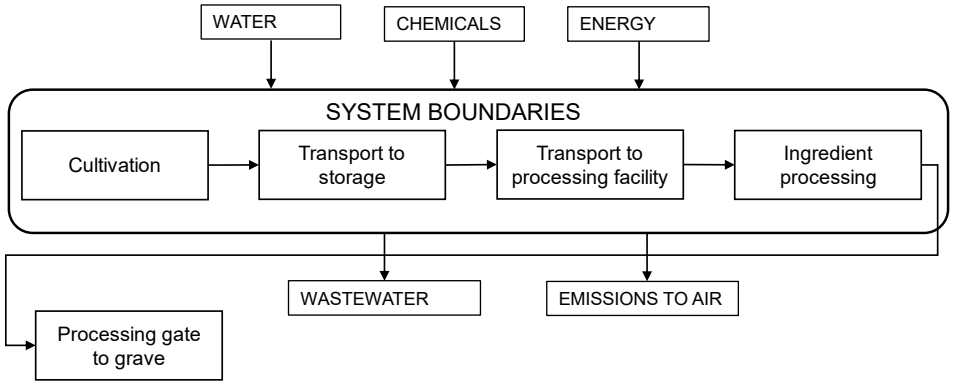

Figure 3.1 System boundaries for the life cycle assessment of the production of food ingredients.

from the Agri-footprint 5.0 database (Table A 3.1 in the Appendix).

3.3.1 Cultivation & transportation

Both the agricultural production and the transportation to storage or feed plants were based on the impact of the consumption mix of the specific crop in the Netherlands and extracted from the Agri-footprint 5.0 database, which is available in the supplementary data (Table A 3.2).[138] The impact of yellow pea was determined from pea in general, as only this was available in the database. The consumption mix for pea originated from Belgium (0.1%), Canada (4.6%), Czech Republic (1.3%), Estonia (0.2%), Finland (23.4%), France (16.9%), Germany (1.7%), Hungary (8.6%), Lithuania (5.1%), the Netherlands (1.7%), Poland (18.4%), Russia (12.0%), Ukraine (4.3%), the United Kingdom (0.7%), and the United States (1.1%). The lupine seeds originated from Germany (20%) and Australia (80%). The distance between the location of cultivation versus the location of processing will influence the overall environmental impact, which is case study dependent.

3.3.2 Ingredient processing

In this study, only the resources related to the production of the ingredients were considered relevant for the comparison between processes with a different degree of refining. Additional impact by for example construction, maintenance, and cleaning of equipment was not included. The information for the conventional and mild aqueous fractionation of yellow was mainly based on literature.[8] A patent was also used for additional information for the conventional process.[131] The conventional fractionation of lupine was also based on literature[17] and a patent.[132] The mild aqueous fractionation of lupine was based on the same study as the conventional variant and a patent.[139] As the exergy study by Berghout et al.[17] mainly focussed on the drying and oil extraction steps, the processes were completed using information from the fractionation of yellow pea in this study. The dry fractionation of both crops was based on literature.[52] The different degrees of processing were defined as follows and presented in more detail including flow diagrams in the supplementary information.

In general, all wet processes start with a milling and steeping step. The initial

separation of starch and fibre fractions is done with either hydrocyclones or decanters. These fractions are subsequently dried by evaporation or a vacuum drum filter and a pneumatic ring drier. In the conventional way of fractionation, the oil (in the case of oil-bearing crops) is initially removed, after which the proteins are precipitated isoelectrically and finally neutralized. The precipitation and neutralization steps are omitted in the mild aqueous fractionation for yellow pea. The protein-rich fractions are concentrated using ultrafiltration and/or dried through spray drying and a fluidized bed. In the milder process for oil-bearing crops the precipitation step is still included, yet the oil extraction is omitted. The mildest fractionation method for both crops is dry fractionation, using a milling step and air classifier. In hybrid fractionation, the fine fraction obtained through dry fractionation is further processed using the mild aqueous fractionation. A detailed description including the assumptions for these fractionation methods can be found in the supplementary information. The energy requirements were based on the parameters presented in Table 3.2.

3.4 Life cycle impact assessment

The life cycle impact (LCI) assessment was performed by combining the existing impact of the agricultural production and transportation and the modelled fractionation processes using Simapro LCA software version 9.0 and the ReCiPe 2016 Midpoint (H) V1.03 method.[140] First, the LCI data points and the connected background processes (Table A 3.1) are combined and compiled into an inventory list of the substance flows (e.g., “methane, biogenic” to Air in kg CH4/kg product or “dinitrogen monoxide” to Air in kg N2O/kg product). Next, the ReCiPe method will translate the substance flows to the various impact categories, based on characterization factors as indicated in the ReCiPe method. A substance flow can influence more impact categories, such as “dinitrogen monoxide” to Air has a characterization factor of 298 kg CO_2 eq/kg N2O for global warming and 0.011 kg CFC-11/kg N2O for Stratospheric Ozone Depletion.

Table 3.2 Energy requirements for the process parameters used in the LCA impact assessment, a detailed description of all processes can be found in the supplementary information.

Process parameters	Unit	Energy consumption
Electricity use:		
Mill	MJ/kg feed	0.5 [52]
Air classifier	MJ/kg feed	0.023 [52]
Dispersion mixing	MJ/kg protein	1.5 [52]
Hydrocyclones	MJ/kg feed	0.0018 [8]
Centrifugal decanter	MJ/kg feed	0.0024 [8]
Vacuum drum filter	MJ/m^3 feed	99 [8]
Ultrafiltration	MJ/m^3 feed	10 [8]
Air pump	MJ/kg feed	0.019 [8]
Electricity oil extractor	MJ/kg flour	0.07 [17]
Fuel energy use drying:		
Pneumatic ring dryer	MJ/kg water removed	4.3 [8]
Evaporator	MJ/kg water removed	0.8 [8]
Spray dryer	MJ/kg water removed	4.8 [52]
Fluidized bed	MJ/kg water removed	6.75 [8]

3.5 Results & discussion

Environmental impacts of the different fractionation methods for processing 1000 kg of yellow pea and lupine with a lower degree of refining are compared excluding cultivation and transportation (Table 3.3). In general, all impact categories decrease with a lower degree of refining in the case of yellow pea and lupine. Both the conventional and mild aqueous fractionation of lupine have a higher environmental impact compared to yellow pea, which can be attributed to the differences between processing starch- or oil-bearing crops. Dry fractionation of lupine seeds is similar to yellow pea as the process is the same. In the following paragraphs, the main differences between the environmental impacts will be discussed and related to the differences among the fractionation processes. The overall impact to process 1000 kg of crop including cultivation and transportation is presented in the Appendix.

Table 3.3 All environmental impacts for the fractionation of yellow pea and lupine to process 1000 kg crop, excluding cultivation. Green-yellow-orange-red highlighted boxes indicate the fractions with the lowest to highest impact among all fractions from both lupine and yellow pea. lupine and yellow pea. In bold are the impact categories that are further discussed.

		Yellow pea				Lupine			
Impact category	Unit	CF	MF	DF	HF	CF	MF	DF	HF
Global warming	kg CO_2 eq	940	541	107	251	1464	1156	107	564
Stratospheric ozone depletion	kg CFC11 eq	0.00020	0.00010	0.00003	0.00006	0.00034	0.00030	0.00003	0.00014
Ionizing radiation	kBq Co-60 eq	16.8	9.0	3.8	6.0	24.5	24.6	3.8	12.1
Ozone formation, Human health	kg NO_x eq	0.012	0.0042	0.0008	0.0019	0.0979	0.0613	0.0008	0.0563
Fine particulate matter formation	kg PM2.5 eq	0.26	0.13	0.03	0.07	0.48	0.41	0.03	0.19
Ozone formation, Terrestrial ecosystems	kg NO_x eq	0.0168	0.0068	0.0013	0.0031	0.1277	0.0688	0.0013	0.0607
Terrestrial acidification	kg SO_2 eq	0.83	0.41	0.10	0.21	1.57	1.33	0.10	0.61
Freshwater eutrophication	kg P eq	0.00006	0.00003	0.00001	0.00001	0.00013	0.00012	0.00001	0.00006
Marine eutrophication	kg N eq	0.0040	0.0016	0.0002	0.0006	0.0091	0.0067	0.0002	0.0027
Terrestrial ecotoxicity	kg 1,4-DCB	39.1	18.3	6.8	11.5	81.1	77.7	6.8	37.2
Freshwater ecotoxicity	kg 1,4-DCB	0.017	0.009	0.003	0.005	0.028	0.026	0.003	0.013
Marine ecotoxicity	kg 1,4-DCB	0.084	0.045	0.013	0.025	0.157	0.143	0.013	0.082
Human carcinogenic toxicity	kg 1,4-DCB	0.28	0.17	0.02	0.07	0.43	0.31	0.02	0.15
Human non-carcinogenic toxicity	kg 1,4-DCB	2.04	1.07	0.35	0.62	4.11	3.32	0.35	1.69
Land use	m^2a crop eq	0.11	0.07	0.00	0.02	0.20	0.14	0.00	0.08
Mineral resource scarcity	kg Cu eq	0.030	0.017	0.004	0.009	0.048	0.041	0.004	0.021
Fossil resource scarcity	kg oil eq	248	153	25	67	370	283	25	146
Water consumption	m^3	15.2	10.4	0.2	3.1	27.0	18.3	0.2	9.6

3.5.1 Comparison impacts from a high to low degree of processing

Four impact categories were selected to compare the impacts of only the different fractionation processes; hence, without cultivation and transportation. The categories represent the impact on 1) ecosystems using global warming potential, 2) human health with human carcinogenic toxicity, and 3) resources using water consumption expressed in blue water use and 4) fossil resource scarcity. For processing 1000 kg yellow pea or lupine, all environmental impact factors decrease drastically compared to the conventional way of isolation, which is set as 100% in Figure 3.2. The results for human carcinogenic toxicity introduce a degree of uncertainty since it is difficult to measure, calculate, and translate into a single impact indicator.[141] Therefore, the numbers are used for a comparison of health damage, rather than absolute values. Mild aqueous fractionation reduces all environmental indicators with approximately 30–40% for yellow pea and 20–35% for lupine compared to conventional fractionation. The impact of yellow pea can be decreased more compared to lupine due to several differences between processing oil- or starch-bearing crops. The fractionation of lupine requires an oil extraction step and involves larger quantities of water compared to the fractionation of yellow pea. This water eventually needs to be removed again to produce dried ingredients. Moreover, in the mild aqueous fractionation of lupine, protein is still extracted by isoelectric precipitation. This step is not used in the milder fractionation of yellow pea. Therefore, the extraction of protein from lupine seeds could be rendered even more sustainable by omitting oil extraction and protein precipitation. Using dry instead of conventional fractionation decreases the impacts by up to 99% for both crops. The impact of hybrid fractionation lays in between the other two methods. Geerts et al.[8] also reported that mild aqueous fractionation has a higher exergy efficiency of 54%, compared to 35% in conventional fractionation, mainly due to the loss of immaterial exergy. Similarly, an exergy efficiency for dry fractionation of 99–100% was found, since all materials and limited amounts of electricity were used. Furthermore, Berghout et al.[17] presented that oil extraction leads to significant exergy losses and destruction. The removal of this step is the main cause of the decrease in environmental indicators in mild aqueous and hybrid fractionation found in this research. The results show that a life cycle assessment can be very

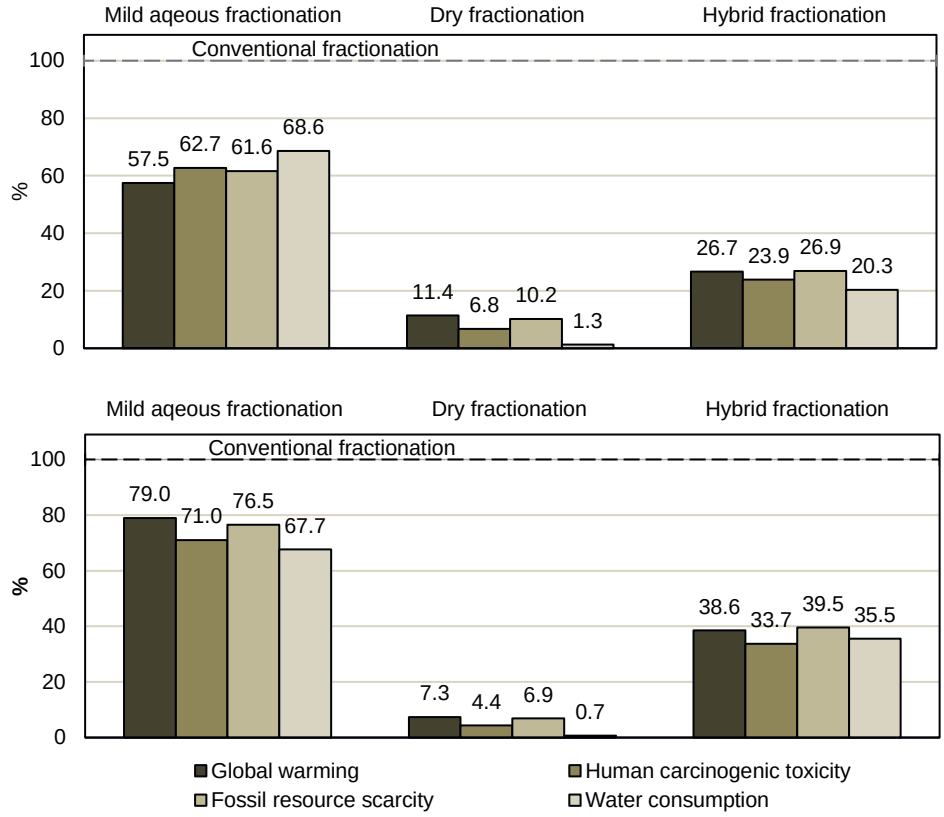

Figure 3.2 Global warming potential (kg CO_2-eq), fossil resource scarcity (kg oil eq), human carcinogenic toxicity (kg 1,4-DCB), and water consumption (m^3) of the mild fractionation of 1000 kg crop without cultivation as a percentage from the conventional fractionation method for yellow pea (top Figure) and lupine (bottom Figure)

valuable in sustainability analyses next to exergy analyses. They provide a detailed insight into the translation of higher efficiencies to environmental impacts.

3.5.2 Environmental impacts broken down per process step

The impact to process 1000 kg of crops without cultivation and transportation is divided into the main processing steps: milling, oil removal, steeping, separation (hydrocylone or decanter), precipitation and neutralization, and drying or ultrafiltration (Figure 3.3 and Figure 3.4). This division aimed at determining the origin of the impact of the fractionation of yellow pea and lupine.

3.5.2.1 Yellow pea

For yellow pea, the global warming potential, human carcinogenicity, fossil resource scarcity, and water consumption all decrease with a lower degree of refining. A closer look into the process shows that the omission of the acidic precipitation step, with the associated heating step, is mostly responsible for the decrease in all impact categories (Figure 3.3). With the removal of this step, less or even no water, no chemicals, and less electricity or process steam were required. In addition, treating the wastewater of the fraction containing soluble solids after precipitation leads to negative water usage, as water is returned to the environment. As the precipitation step is removed, the subsequent decanter step could also be left out. The latter was mainly responsible for the impact in the separation category and created a large waste stream. The impact of the hydrocyclones/decanters category is also lower for mild aqueous fractionation since starch is separated in decanters. In contrast, conventional fractionation uses hydrocyclones, which require more energy compared to decanters. However, the decanters in mild aqueous fractionation produce a larger starch fraction compared to the conventional method. This requires more energy to be dried by the subsequent pneumatic drying step. Therefore, replacing decanters with hydrocyclones in this case study will have a low effect on the overall environmental impact. The environmental impact is further reduced with dry fractionation, which results in a decrease ranging from 87 to 99% in impact. Overall, drying is responsible for a large part of the impact and does not decrease very dramatically in the mild aqueous fractionation methods. Geerts et al.[8] showed that drying was the main driver for exergy losses in both conventional and mild wet fractionation, whereas the precipitation step was responsible for only 10% of the exergy losses. This is most likely a slight underestimation because the material and immaterial costs of the materials (acids and bases) were not considered in that study. This can now also be translated into the environmental impact, which shows that the omission of the extraction-precipitation process can drastically reduce the environmental indicators.

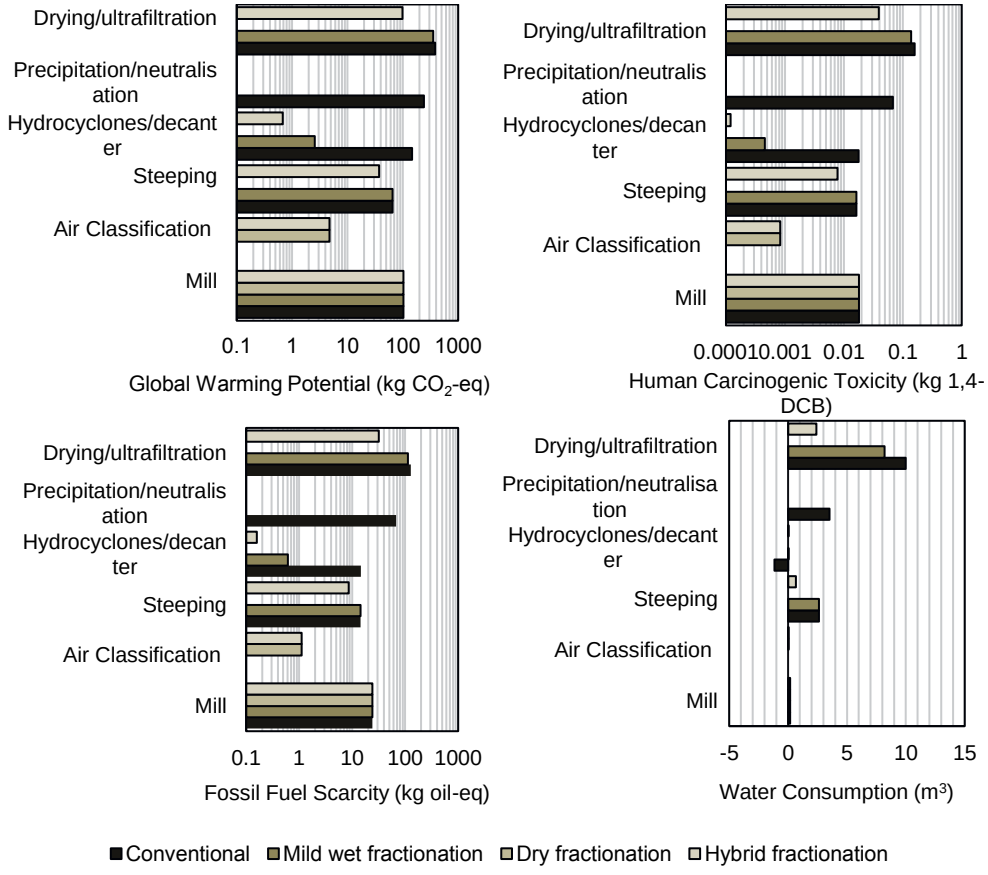

Figure 3.3 The global warming potential, human carcinogenic toxicity, fossil fuel scarcity, and water consumption for each process step in conventional, mild aqueous, dry, and hybrid fractionation of yellow pea. Please note the log scale.

3.5.2.2 Lupine

Fractionating lupine without the removal of oil reduces all impact categories, as the use of hexane and the distillation of hexane is omitted. Through mild aqueous fractionation of lupine, the oil fraction (approximately 10% in lupine flour based on dry matter) ends up in very small amounts in the protein isolate (0.02–0.07 g oil/g protein isolate) and the rest in the fibre-rich pellet. The functional properties of the protein-rich fraction are not significantly altered by the presence of oil compared to the conventional protein isolate.[142] The same was found for the dry fractionated fine fraction of lupine, which showed increased foam stability compared to the

conventional lupine protein isolates.[48] For both crops, dry fractionation leads to the largest decrease in footprint, as less electricity, no process steam, and no chemicals are used; of course, at the cost of product purity. The combined method is situated between the dry and mild aqueous fractionation as the fine fraction was further processed (Figure 3.4). The drying steps are responsible for a large part of the footprint. This is corroborated by Berghout et al. [17] who also considered alternatives for processes that require a lot of energy for drying, such as dry fractionation. The water use was reduced in the milder fractionation method, mainly due to the different ratio of water to material that was used. As for fractionation yellow pea, wastewater was also returned to the environment after receiving treatments. Especially the human carcinogenic toxicity is reduced drastically with the removal of the oil step. This is mostly attributed to heating hexane during the oil distillation rather than the use of hexane, since the hexane is reused almost completely. More specifically, only 3 kg hexane per ton lupine protein isolate is typically lost into the atmosphere during production[143] and hence, affects the human carcinogenic toxicity, whereas the remainder is not considered.

In general, these results confirm the importance of reducing the degree of processing for food ingredients from the viewpoint of the environmental impact. Mainly the removal of the alkaline-acidic extraction-precipitation process and the oil extraction are responsible for this. Moreover, drying remains a dominant process step, which indicates that this remains an important issue to focus on. One should bear in mind that the milder fractionation pathways are still based on lab-scale processes that could be better optimized when developed for industrial scale. Therefore, the results may still change due to upcoming developments. We do believe however that the conclusions will not change, and in fact, will only become more distinct due to better efficiencies on larger scales. Furthermore, the different processes deliver products with very different qualities and properties. The conventional protein isolates are quite pure, while the concentrates from the other processes contain significant amounts of other components, but may still have good, though different functionality. Therefore, the direct comparison of these processes is not without complexity.

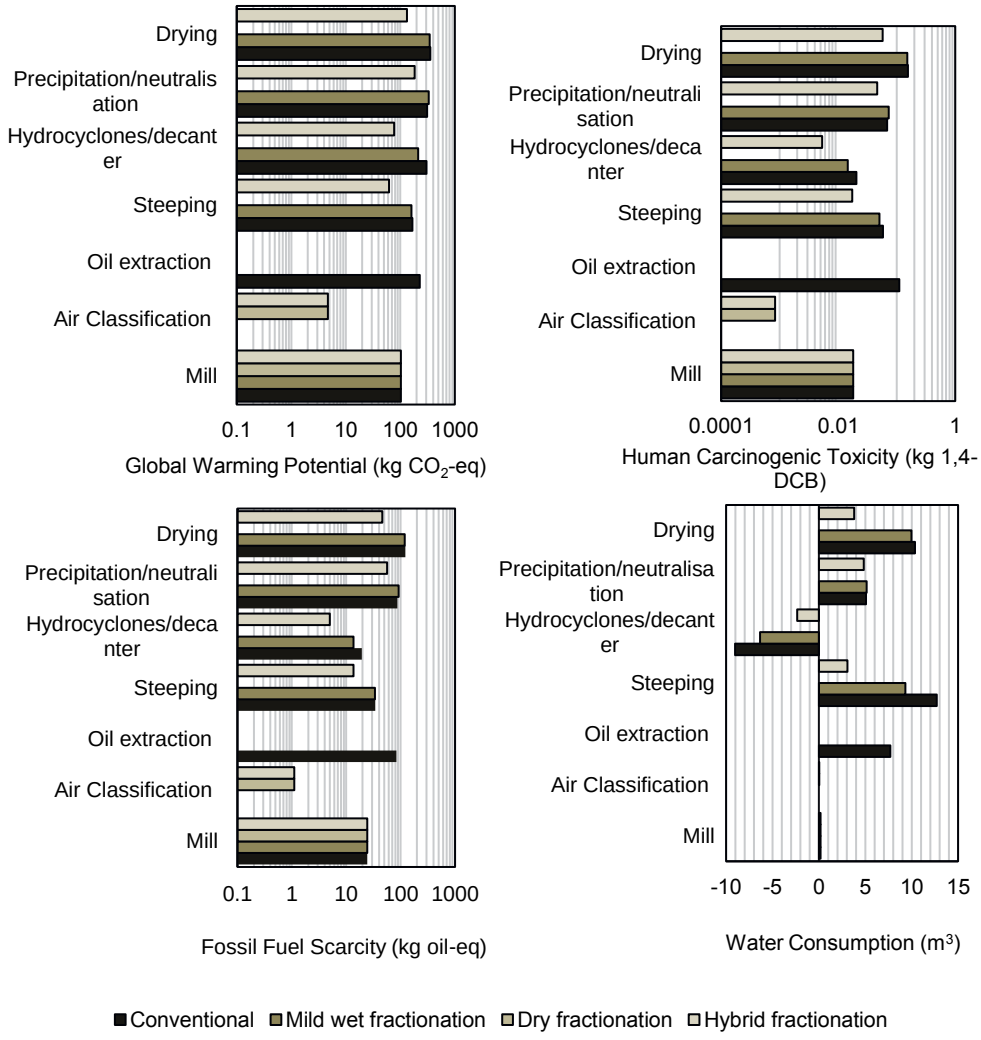

Figure 3.4 The global warming potential, human carcinogenic toxicity, fossil fuel scarcity, and water consumption for each process step in conventional, mild aqueous, dry, and hybrid fractionation of lupine. Please note the log scale.

3.5.3 Comparison between cultivation and processing

Until now the environmental impact of different ways of refining 1000 kg yellow pea and lupine was discussed without including cultivation and transportation. Figure 3.5 illustrates the effect of including the global warming potential, human carcinogenic toxicity, fossil resource scarcity, and water consumption of cultivation, transportation, and processing of 1000 kg of lupine or yellow pea. The combined

environmental impact to produce 1 kg of ingredient and protein can be found in the Appendix. The global warming potential and fossil resource scarcity of the cultivation of lupine or yellow pea are similar. The cultivation of lupine has more effect on human carcinogenic toxicity compared to yellow pea. The latter is explained by the higher toxicity emission related to the use of insecticides and herbicides outside of the EU (lupine originates partly from Australia) and longer transportation distances. In contrast, yellow pea requires more water during cultivation compared to lupine, as it is cultivated using irrigation.[138]

For the global warming potential and fossil resource scarcity of both conventional and mild wet fractionation, the processing is dominant over the cultivation due to

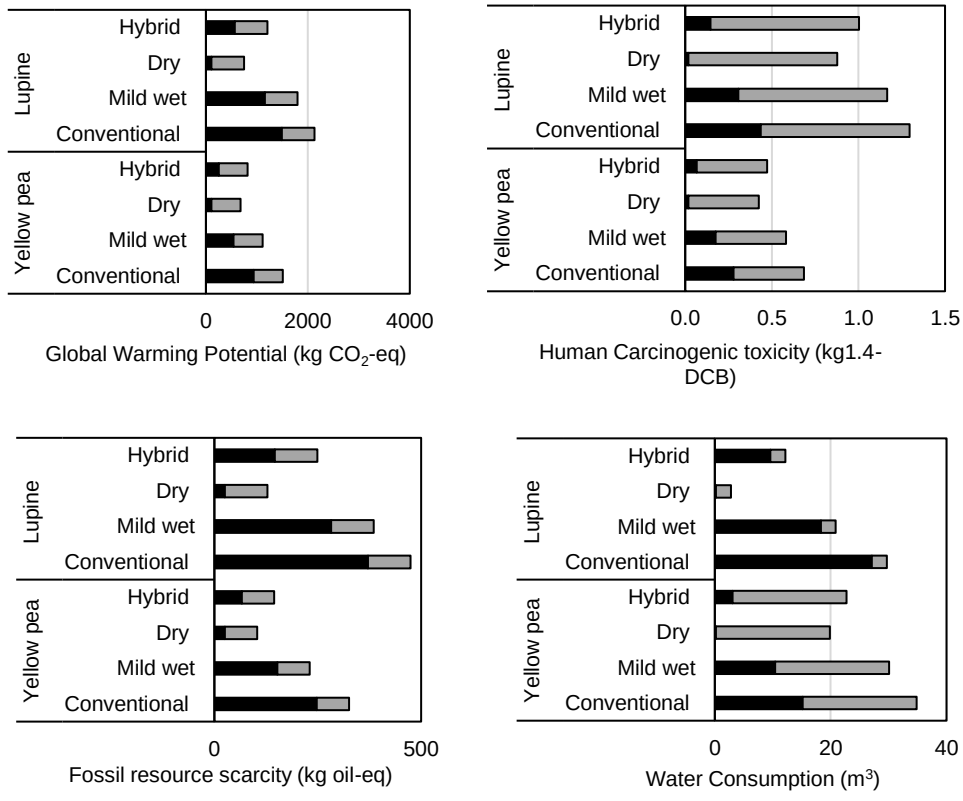

Figure 3.5 The global warming potential, human carcinogenic toxicity, fossil resource scarcity, and water consumption of cultivation (grey) and processing (black) for yellow pea and lupine for processing 1000 kg crop.

the relatively low energy requirements of the cultivation. The human carcinogenic toxicity of cultivation is dominant for all processes, mainly caused by the use of pesticides. With regards to water consumption, the processing of lupine contributes substantially more than its cultivation. This is different for yellow pea due to the difference in irrigation and water use in the steeping step during fractionation. The ratio between cultivation and fractionation is different for each environmental indicator (full dataset in the Appendix). As both have a significant contribution, it is important to include both the environmental impact of cultivation and fractionation when selecting the most sustainable ingredients. Evidently, the place and method of cultivation highly influence the environmental indicators as different countries have different climates and distances to the processing location. This could result in a different outcome in a case study with different cultivation locations. Nevertheless, the comparisons made in this case study are to emphasize the impact of cultivation compared to processing.

3.5.4 Environmental impact of the production of 1 kg of protein

Earlier, it was shown that a lower degree of refining decreases the environmental impact of the processes and hence, also of the embedded environmental impacts in the food ingredients. Moreover, we showed in the previous section that both the impact of cultivation, transportation, and processing of the crops should be included in any sustainability assessment. One should however bear in mind that a lower degree of processing results in fractions with a lower protein purity, which is considered a key component in food ingredients. Therefore, instead of using the amount of processed raw material as a functional unit, the global warming potential of the fractionation processes can also be expressed per kg of protein within that ingredient, now including the impact of cultivation and transportation (Figure 3.6). The other environmental impact indicators expressed per kg of protein are presented in the Appendix and follow the same reasoning as the global warming potential. The conversion from total mass to protein as a functional unit is inversely related to the environmental indicators due to the low purity of the milder fractionated ingredients. Nevertheless, a lower degree of refining can still deliver the same amount of protein with a lower footprint (Figure 3.6). The carbon footprint to produce conventional yellow pea protein isolate is 5.3 kg CO_2-equivalents/kg protein

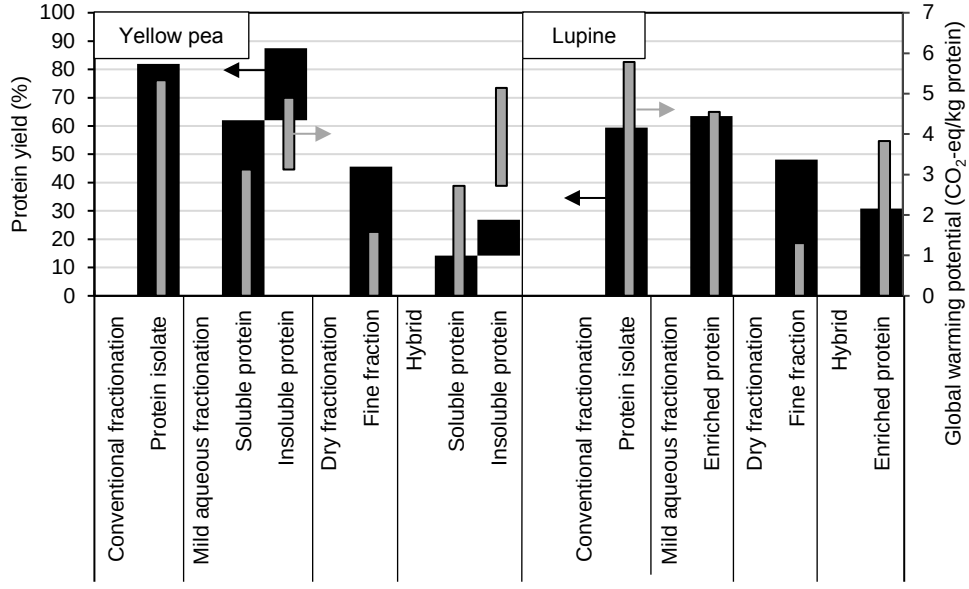

Figure 3.6 Protein yield (protein in the ingredient as a percentage of the protein in the crop) (left axis in black) and the global warming potential (right axis in grey) to produce 1 kg of protein in each fraction.

and 4.9 kg CO_2-equivalents/kg protein for mild aqueous fractions. The fine fraction has the lowest climate change potential with 1.6 kg CO_2-equivalents/kg protein. The global warming potential of hybrid protein-rich fractions is 5.1 kg CO_2-equivalents/kg protein. The impact of the hybrid protein fractions surpasses the impact of the mild aqueous protein fractions. This is due to a matter of impact allocation to the protein/fibre-rich stream and the starch stream in the first decanter. As the starch-rich fraction produced from the fine fraction is smaller than the starch-rich fraction that originates from yellow pea flour, relatively more impact is allocated to the hybrid protein-rich fractions. Moreover, the global warming potential of the protein-rich fractions from lupine is slightly higher compared to yellow pea. The conventional fractionation of lupine requires oil removal and more water compared to the fractionation of yellow pea. The excess water also needs to be evaporated, which requires additional energy. The climate change potential for the conventional protein isolate of lupine is 5.8 kg CO_2-equivalents/kg protein, while the milder aqueous variant is 4.6 kg CO_2-equivalents/kg protein. The dry fractionation again

has the lowest carbon footprint, of 1.3 kg CO_2-equivalents while the hybrid method is 3.8 kg CO_2-equivalents per kg protein. This proves that the mildly fractionated ingredients have the potential to deliver as much protein as the conventional way, but with a reduced carbon footprint. The milder fractions also have a lower carbon footprint than for example soy protein isolate, which was estimated by Thrane et al.[144] 6.1 kg CO_2-equivalents/kg protein.

Comparing the impact between different processes and fractions, one should realize that the impact per fraction (e.g. starch isolate and protein isolate) is dependent on the allocation of the process steps and the impact of cultivation. More specifically, for conventional fractionation, the impact of cultivation is allocated according to the yield of the three fractions; the protein-, fibre-, and oil/starch-rich fractions. For dry fractionation the impact is allocated to only the fine and coarse fractions, meaning a relative higher cultivation impact per fraction. It is however still debatable whether the fibre fraction from conventional fractionation is considered to be food grade; hence, a useful fraction. The removal of such a fraction from the allocation will increase the final impact of each fraction. Therefore, the allocation to valuable fractions should always be considered carefully, either based on mass or economical value.

Next to the reduction of the global warming potential during the production of food ingredients, it is evident that for a sustainable food chain a minimal loss of valuable components is essential. Even though conventional fractionation of yellow pea delivers a high protein yield of 82% (Figure 3.6), all other protein originally present is lost. Even if they could be recovered, they are not food grade anymore. In contrast, the protein-rich ingredients from mild aqueous fractionation of yellow pea combined have a higher protein yield of 87%, assuming that both the insoluble and the soluble proteins are recovered. The rest of the proteins end up in the starch fraction and could still be used for human consumption. A similar relation is found for the fractionation of lupine. The protein yield of mild aqueous fractionation of lupine is 64%, compared to conventional fractionation of 59%, with a lower impact. These findings indicate that besides the lower global warming potential of the milder alternatives to conventional fractionation, they also deliver a higher protein yield.

Dry fractionation features a protein yield of 100% for both fractions. It is assumed that no fractions are degraded because the milled flour is just separated into two fractions through an air classifier. On top of that, the global warming potential is very low for this method. However, the downside of this method is the low purity that is obtained; with a significant amount of protein ending up in the protein lean coarse fraction. A method to increase the purity was proposed by Berghout et al. [17], by further purifying the fine fraction using mild aqueous fractionation. Interestingly, combining dry fractionation and mild aqueous fractionation for yellow pea only slightly changes the purity of the fine fraction from 43% to 62 and 37% for the soluble and insoluble protein fraction respectively. As the fine fraction is the starting material of the second step of the hybrid fractionation, the protein yield compared to the initial crop is approximately 15 and 13% for the soluble and insoluble protein fraction, respectively. Next to the rather slight increase in purity with a low total protein yield of the hybrid fractions, the hybrid fractionation comes at a cost of 3.5 CO_2-equivalents/kg of protein extra. One should however realize that no protein is lost during both mild fractionation techniques, since the proteins in the coarse fraction can be utilized completely in food, as opposed to conventional fractionation. In addition, the combined method is still lab-based and can be optimized. The hybrid fractionation of lupine increases the purity of the fine fraction from 58 to 87%, with a protein yield of 30%.[17] While this looks very promising, the values are uncertain since it is calculated theoretically by the author. The same distribution of mild aqueous fractionation was used, starting with lupine flour rather than the fine fraction. This might be an unjustified assumption as the fine fraction has a different composition; hence, the hybrid protein-rich fraction will have a different composition. This could also be an explanation for why the carbon footprint of the protein-rich hybrid fraction does not surpass the mild aqueous protein-rich fraction as is the case with yellow pea. Therefore, we conclude for now that for yellow pea, hybrid fractionation can slightly increase the purity of the fine fraction after dry fractionation, yet, this comes at a high cost in terms of global warming potential. For lupine, the situation may be the opposite, but more research is required. Overall, the use of protein as a functional unit in this sustainability assessment indicates that a lower degree of refining can deliver as much protein as the conventional

fractionation, with a lower footprint. On top of that, milder methods have a higher protein yield, in which close to no protein is lost.

3.6 Conclusion

This study quantified the environmental impact of several degrees of refining for food ingredients. The environmental impact of the production of plant protein-rich ingredients can be reduced substantially by decreasing the degree of refining, which we assessed for yellow pea and lupine seeds. Overall, processing is a very important element in the total environmental impact and in some cases larger than that of the crop cultivation. Therefore, any assessment should include both. The omission of the protein precipitation step in mild aqueous fractionation significantly decreases the global warming potential, human carcinogenic toxicity, water consumption, and fossil resource scarcity by 30–40%. Moreover, the removal of the oil extraction step in the fractionation process of oil-rich seeds such as lupine decreases these impact categories by 20–30%. The final drying steps remain mostly responsible for the impact in all categories. The largest decrease in global warming potential was obtained with only dry processing with a decrease of up to 93% compared to the conventional way of fractionation. With the amount of recovered protein as the functional unit, the processes that refine less still have a lower impact compared to the conventional isolation method. From all ingredients, the conventional method gives the highest protein yield per protein-rich fraction, although at the costs of a higher environmental impact. In general, the consideration of purity, yield, and environmental indicators offer more insight into the process of choosing the right ingredients, which will benefit the sustainability of the food chain.

3.7 Appendix

3.7.1 Background impacts life cycle assessment

Table A 3.1 Overview of the background life cycle impacts used for the life cycle assessment based on Agri-footprint® 5.0.[138]

Category	Impact name
Electricity	Electricity mix, AC, consumption mix, at consumer, < 1kV NL S System - Copied from ELCD
Steam	Process steam from natural gas, heat plant, consumption mix, at plant, MJ, NL S System - Copied from ELCD
Water	Drinking water, water purification treatment, production mix, at plant, from surface water RER S System - Copied from ELCD
Sodium Hydroxide	Sodium hydroxide (50% NaOH) (mix), at plant/RER Mass
Waste	Wastewater treatment, domestic wastewater according to the Directive 91/271/EEC concerning urban wastewater treatment, at wastewater treatment plant EU-27 S System - Copied from ELCD
Hydrochloric Acid	Hydrochloric acid (30% HCl) (Mannheim), at plant/RER Mass
Hexane	Hexane emission to Air (unspecified)

Table A 3.2 Weighted averages of the life cycle inventory regarding the cultivation of consumption mix in The Netherlands of yellow pea and lupine obtained from the Agri-footprint 5.0 database.[138] "n" is the number of countries included in the consumption mix: lupine: (20%) and Australia (80%) and yellow pea: Belgium (0.1%), Canada (4.6%), Czech Republic (1.3%), Estonia (0.2%), Finland (23.4%), France (16.9%), Germany (1.7%), Hungary (8.6%), Lithuania (5.1%), the Netherlands (1.7%), Poland (18.4%), Russia (12.0%), Ukraine (4.3%), the United Kingdom (0.7%), and the United States (1.1%).

Category	Lupine (n = 2)	Yellow pea (n = 15)
Yield (kg/ha)	1383.0	2632.3
straw (kg/ha)	252.6	395.0
Manure N (kg N/ha)	8.8	16.3
Manure P (kg P_2O_5/ha)	5.3	9.8
Inorganic Fertilizers N (kg N/ha)	4.0	22.6
Inorganic Fertilizers P (kg P_2O_5/ha)	32.0	30.2
Inorganic Fertilizers K (kg K_2O/ha)	12.0	39.9
Insecticide (kg/ha)	0.5	0.1
Fungicide (kg/ha)	0.5	0.8
Herbicide (kg/ha)	4.8	3.8
Diesel (MJ/ha)	2336.0	3481.3
Electricity (MJ/ha)	0.0	16.5
Blue Water (m^3/ha)	0.0	44.2
Start material (kg/ha)	62.8	139.8
Transport (km/ha)	57.7	87.1

3.7.2 Description of fractionation processes

Figure A 3.1, Figure A 3.2, and Figure A 3.3 provide a detailed description of the conventional (CF), mild aqueous (MAF), and dry fractionation (DF) respectively of yellow pea and lupine. In this research, the combined method of mild aqueous and dry fractionation, or hybrid fractionation was also considered (HF), in which the fine fraction is used as the starting material for the mild aqueous processes for both lupine and yellow pea. In the following paragraphs, the resource inputs of all processes are described, which are processed in SimaPro with the background impacts listed in Table A 3.1.

CF and MAF of yellow pea are mostly adapted from the proposed method from Geerts et al.[8] and a patent from Roquette, which is a plant based ingredient producer.[131] For both processes, 1000 kg of yellow peas are initially milled and soaked in a 25%wt solution, which requires approximately 0.5 MJ/kg feed electricity and 1.5 MJ/kg protein electricity. During soaking, 3% of the yellow pea solution is lost and considered waste. Subsequently, in CF, starch isolates are produced by a separation step of a series of 14 hydrocyclones that require 1.8 kJ/kg feed per hydrocylone electricity and approximately 0.5 kg water per kg feed. The MAF produces an enriched starch fraction using a decanter instead, to separate the starch, which requires 2.4 kJ/kg feed. Both starch fractions are initially concentrated to 60% dry matter using a vacuum drum filter, with an electricity use of 99 MJ/m^3 feed, after which the moisture is reduced to 12.4% dry matter in a pneumatic ring drier. The latter requires energy, in the form of steam, of approximately 4.3 MJ/kg water removed and 5% water of the incoming feed to clean the equipment. In CF, the yellow pea solution without starch is separated another time using a centrifugal decanter to remove the fibre fraction, which is dried using an evaporator to approximately 17% dry matter. This process requires approximately 0.8 MJ/kg water removed. It should be noted that the dry matter from the fibre fraction is different from the proposed by Geerts et al.[8] (6% dry matter), the main reason is that the evaporated is not commonly used to dry to such a low percentage of dry matter. Therefore, the dry matter of evaporated potato fibre was used in this case.[122]

Next, the remaining mixture containing proteins and soluble materials is subjected to isoelectric and centrifugal separation for CF and MAF respectively. In isoelectric precipitation, the proteins are flocculated by reducing the pH to 4.5 and heating to 60°C using hydrochloric acid, water, mixing electricity, and process steam. The proteins are subsequently further matured in a tank, requiring mixing electricity. The soluble materials are now separated from the precipitated proteins and considered waste. The proteins are neutralized using sodium hydroxide, water, and mixing electricity. The remaining solution for MAF is separated by a centrifugal decanter, after which the pellet containing the insoluble proteins is evaporated to 17% dry matter (using the same reasoning as above). The supernatant containing the soluble solids and proteins is concentrated using ultrafiltration, which requires 10 MJ/m^3 electricity, and produces the retentate, which is considered waste. Both the final protein isolate and the soluble protein from CF and MAF respectively are dried using a spray drier to obtain a dry matter of 20%, requiring 4.8 MJ/kg evaporated water of process steam. Both fractions are further dried to approximately 10% dry matter using a fluidized bed, which needs another 6.75 MJ/kg removed water of process steam. After all drying steps, the ingredients are cooled down using air, required electricity of an air pump of 1.9 kJ/kg feed.

The CF and MAF of lupine seeds are similar to yellow pea. As there was no detailed process available yet, a specific process is composed of information from literature[17], a patent[132], and assumptions. When no information could be obtained from these sources, the fractionation of yellow pea was used. The main difference is that CF includes an additional oil removal step after the milling step in which hexane is added to the milled lupine seeds with a sample to solvent ratio of 1:2 (on an industrial scale) and evaporated at 40°C. Berghout et al.[17] indicated that the amount of hexane differs among lab-, pilot-, or industrial-scale. As this process in theory for the industry, that ratio is taken. For the extraction process steam is required to heat the sample, 0.07 MJ/kg flour electricity is needed, and cooling water is used in approximately 1:1 ratio. During the extraction, approximately 2 g of hexane is lost as vapor. After the extraction, the flour is dispersed into the water with a sample to solvent ratio of 1:15. The pH of lupine suspensions is set to 9 using sodium hydroxide, for which mixing electricity is also required. The fibre is subsequently separated from

the mixture in a centrifugal decanter, for which electricity is required. The fibre-rich fraction is evaporated to 17% dry matter using steam.

The remaining dispersion containing proteins and other soluble materials is subjected to a precipitation step, in which the pH is adjusted to 4.5 with hydrochloric acid at 20°C. Subsequently, the solution is matured in a maturation tank. For these steps mixing electricity is required. After this, the protein is neutralized using sodium hydroxide, which required mixing electricity. Some proposals of lupine extraction include a washing step instead of a neutralization step in literature[17] and in practice.[139] However, washing steps require large quantities of water (e.g. in the order of 10,000 kg) and are highly dependent on the size of the feed stream in that step. It seems that for MAF there is a larger feed because the oil is not removed. This will lead to much more water that has to be evaporated. As this process is still not optimized, it was decided to include a neutralization step similar to the fractionation of yellow pea, which is also used in practice. After the neutralization, the protein fraction is initially dried to 20% using a spray drier, after which it is dried to the final dry matter of approximately 10% using a fluidized bed, for which both steam is required. The fluidized bed was introduced to create an equal drying step as in the fractionation of yellow pea. The MAF for lupine seeds is the same as for CF, except for that the complete oil extraction step is omitted and that the full-fat flour is dispersed into the water in a ratio of 1:10, which is also used in industry.[139] A lower ratio is picked here for the same reasoning as above; it will result in unbalanced energy requirements for the drying steps, on which the focus of this research is not.

The dry fractionation for yellow pea and lupin is the same and based on literature [52], in which the crops are milled using the electricity requirements. After the milling, the flours are separated in an air-classifier, which has an electricity requirement of 2.3 kJ/kg feed. No losses are assumed in this process, as the losses of these steps are minimal on the industrial scale.

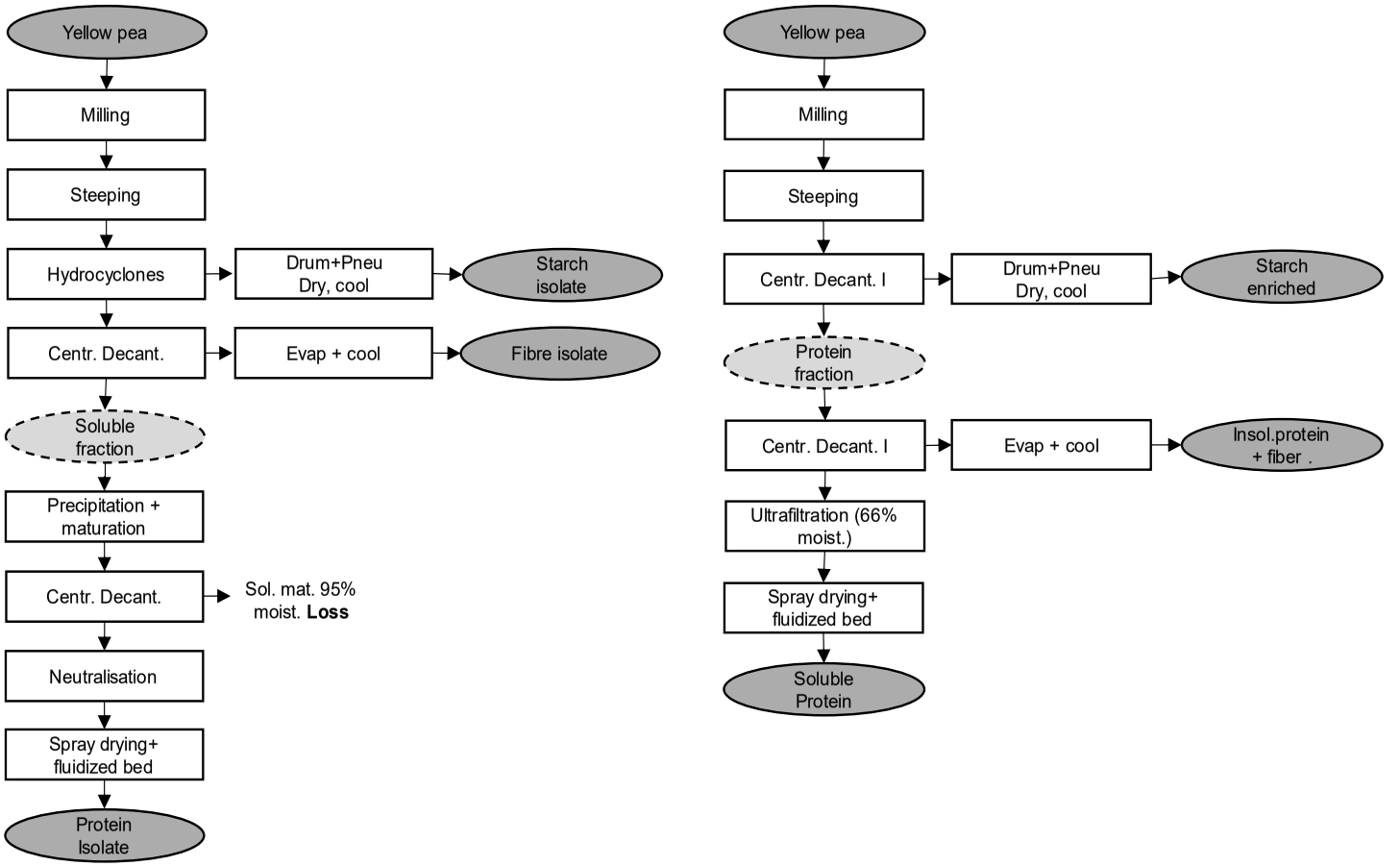

Figure A 3.1 Flow chart of conventional wet fractionation (left) and mild wet fractionation (right) of yellow pea.

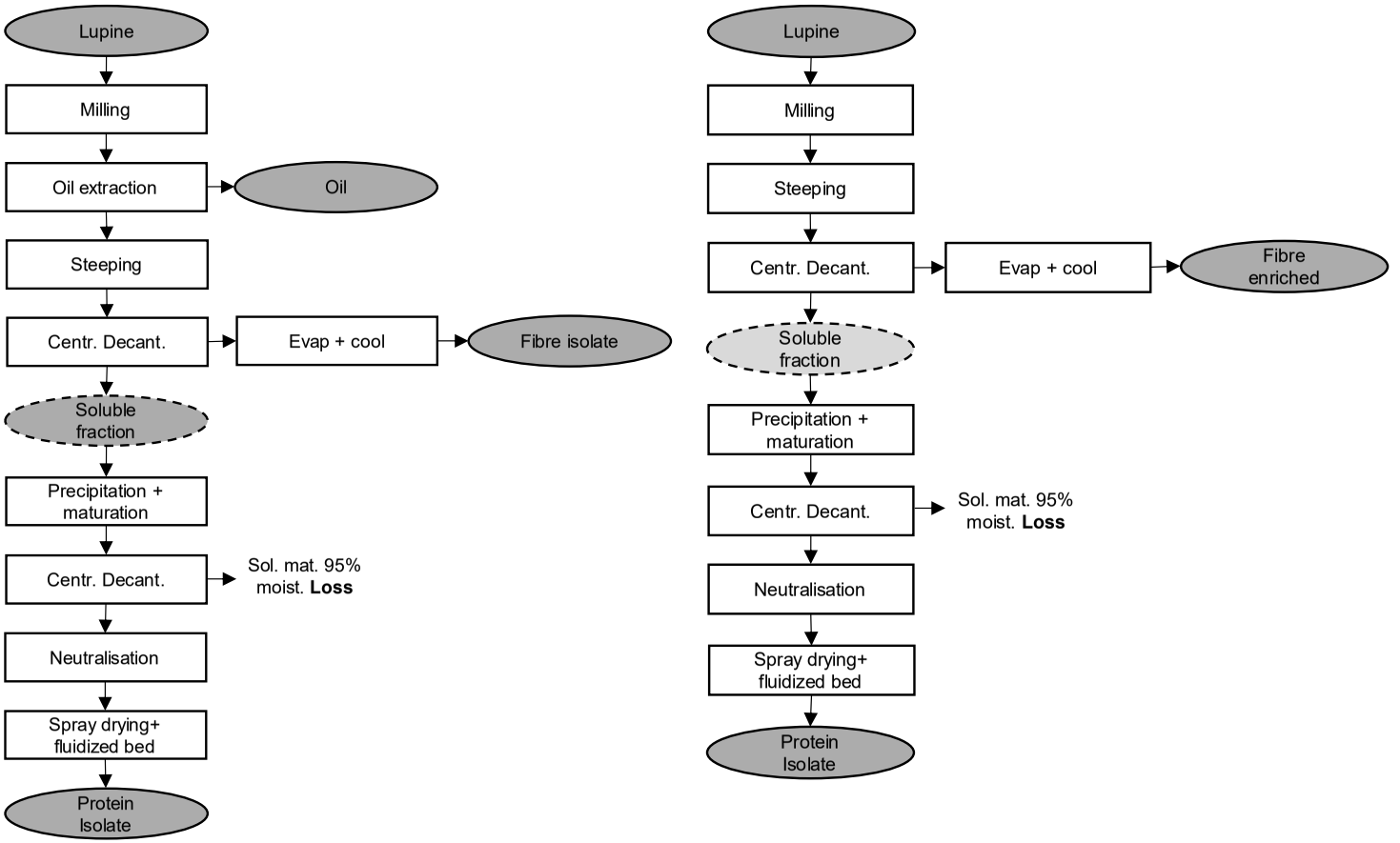

Figure A 3.2 Flow chart of conventional wet fractionation (left) and mild wet fractionation (right) of lupine.

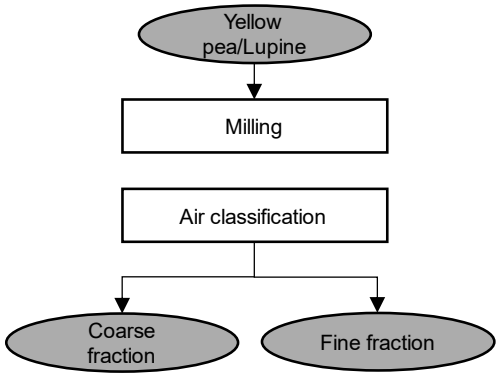

Figure A 3.3 Flow chart of dry fractionation of yellow pea and lupine.

3.7.3 Results for all environmental impact categories.

Table A 3.3 Environmental impacts for the fractionation of yellow pea and lupine to process 1000 kg crop including cultivation.

		Yellow Pea				Lupine			
Impact category	Unit	CF	MF	CF	HF	CF	MF	CF	HF
Global warming	kg CO_2 eq	1509	1101	676	820	2106	1798	750	1207
Stratospheric ozone depletion	kg CFC11 eq	7.8E-03	7.6E-03	7.6E-03	7.6E-03	5.1E-03	5.0E-03	4.8E-03	4.9E-03
Ionizing radiation	kBq Co-60 eq	19.9	12.0	6.9	9.1	29.4	29.5	8.7	17.0
Ozone formation, Human health	kg NO_x eq	1.63	1.59	1.61	1.62	3.01	2.97	2.91	2.97
Fine particulate matter formation	kg PM2.5 eq	1.12	0.97	0.89	0.92	1.53	1.45	1.07	1.23
Ozone formation, Terrestrial ecosystems	kg NO_x eq	1.64	1.61	1.63	1.63	3.06	3.00	2.93	2.99
Terrestrial acidification	kg SO_2 eq	6.03	5.53	5.30	5.41	6.64	6.40	5.17	5.68
Freshwater eutrophication	kg P eq	0.38	0.37	0.38	0.38	0.54	0.54	0.54	0.54
Marine eutrophication	kg N eq	2.54	2.50	2.54	2.54	1.91	1.90	1.90	1.90
Terrestrial ecotoxicity	kg 1,4-DCB	31679	31176	31647	31652	42646	42643	42572	42602
Freshwater ecotoxicity	kg 1,4-DCB	55.96	55.10	55.95	55.95	149.35	149.35	149.33	149.34
Marine ecotoxicity	kg 1,4-DCB	102.47	100.87	102.40	102.42	414.70	414.69	414.56	414.63
Human carcinogenic toxicity	kg 1,4-DCB	0.69	0.58	0.43	0.47	1.29	1.17	0.88	1.00
Human non-carcinogenic toxicity	kg 1,4-DCB	3277.5	3226.6	3275.8	3276.1	3108.4	3107.6	3104.7	3106.0
Land use	m^2a crop eq	3517.0	3463.3	3516.8	3516.9	6464.1	6464.0	6463.9	6463.9
Mineral resource scarcity	kg Cu eq	1.08	1.05	1.05	1.06	1.74	1.74	1.70	1.72
Fossil resource scarcity	kg oil eq	326	230	104	145	472	386	128	249
Water consumption	m^3	34.8	29.8	19.8	22.7	29.6	20.9	2.8	12.2

Table A 3.4 Environmental impact for fractions obtained from yellow pea and lupine to produce 1 kg of fraction without cultivation included.

Impact category	Unit	Yellow Pea				Lupine					
		CF	MF	DF	HF	CF	MF	DF	HF		
		Protein Isolate	Insoluble Protein	Soluble Protein	Fine fraction	Insoluble Protein	Soluble Protein	Protein Isolate	Enriched Protein	Fine fraction	Enriched protein
Global warming	kg CO_2 eq	3.16E+00	4.56E-01	1.04E+00	1.07E-01	3.64E-01	1.11E+00	3.94E+00	2.94E+00	1.07E-01	2.73E+00
Stratospheric ozone depletion	kg CFC11 eq	6.87E-07	6.36E-08	1.64E-07	3.18E-08	8.32E-08	1.83E-07	1.06E-06	7.92E-07	3.18E-08	6.78E-07
Ionizing radiation	kBq Co-60 eq	5.21E-02	6.44E-03	1.27E-02	3.79E-03	9.33E-03	1.57E-02	6.55E-02	5.83E-02	3.79E-03	4.93E-02
Ozone formation, Human health	kg NO_x eq	4.66E-05	3.65E-06	8.24E-06	8.04E-07	2.81E-06	8.78E-06	2.86E-04	1.93E-04	8.04E-07	3.75E-04
Fine particulate matter formation	kg PM2.5 eq	8.87E-04	9.12E-05	2.28E-04	3.21E-05	9.37E-05	2.46E-04	1.41E-03	1.06E-03	3.21E-05	9.70E-04
Ozone formation, Terrestrial ecosystems	kg NO_x eq	6.25E-05	5.88E-06	1.33E-05	1.30E-06	4.53E-06	1.41E-05	3.47E-04	2.12E-04	1.30E-06	3.99E-04
Terrestrial acidification	kg SO_2 eq	2.86E-03	2.90E-04	7.31E-04	1.02E-04	2.98E-04	7.88E-04	4.57E-03	3.42E-03	1.02E-04	3.13E-03
Freshwater eutrophication	kg P eq	2.01E-07	1.62E-08	4.07E-08	8.40E-09	2.17E-08	4.58E-08	3.92E-07	3.10E-07	8.40E-09	3.31E-07
Marine eutrophication	kg N eq	1.52E-05	5.30E-07	2.97E-06	1.71E-07	6.34E-07	2.77E-06	3.14E-05	2.00E-05	1.71E-07	1.60E-05
Terrestrial ecotoxicity	kg 1,4-DCB	1.30E-01	1.24E-02	2.73E-02	6.85E-03	1.73E-02	3.21E-02	2.25E-01	1.87E-01	6.85E-03	1.85E-01
Freshwater ecotoxicity	kg 1,4-DCB	5.47E-05	7.08E-06	1.47E-05	3.03E-06	8.38E-06	1.69E-05	7.69E-05	6.36E-05	3.03E-06	6.15E-05
Marine ecotoxicity	kg 1,4-DCB	2.76E-04	3.56E-05	7.61E-05	1.30E-05	3.73E-05	8.56E-05	4.45E-04	3.66E-04	1.30E-05	4.33E-04
Human carcinogenic toxicity	kg 1,4-DCB	9.32E-04	1.73E-04	3.73E-04	1.89E-05	9.99E-05	3.86E-04	9.91E-04	7.19E-04	1.89E-05	7.12E-04
Human non-carcinogenic toxicity	kg 1,4-DCB	6.67E-03	8.08E-04	1.74E-03	3.49E-04	9.40E-04	1.99E-03	1.08E-02	8.18E-03	3.49E-04	8.17E-03
Land use	m^2a crop eq	3.91E-04	8.05E-05	1.77E-04	0.00E+00	2.84E-05	1.77E-04	4.87E-04	3.50E-04	0.00E+00	5.18E-04
Mineral resource scarcity	kg Cu eq	9.53E-05	1.37E-05	2.88E-05	4.40E-06	1.36E-05	3.19E-05	1.28E-04	1.01E-04	4.40E-06	1.01E-04
Fossil resource scarcity	kg oil eq	8.25E-01	1.44E-01	3.10E-01	2.54E-02	9.81E-02	3.29E-01	9.03E-01	6.93E-01	2.54E-02	7.12E-01
Water consumption	m^3	4.32E-02	1.24E-02	2.20E-02	1.93E-04	6.28E-03	2.24E-02	4.31E-02	3.24E-02	1.93E-04	4.54E-02

Table A 3.5 Environmental impact for fractions obtained from yellow pea and lupine to produce 1 kg of fraction with cultivation included.

Impact category	Unit	Yellow Pea						Lupine			
		CF	MF		CF	HF		CF	MF	DF	HF
		Protein Isolate	Insoluble Protein	Soluble Protein	Fine fraction	Insoluble Protein	Soluble Protein	Protein Isolate	Enriched Protein	Fine fraction	Enriched protein
Global warming	kg CO_2 eq	4.20E+00	9.89E-01	1.66E+00	6.76E-01	9.05E-01	1.69E+00	5.02E+00	3.94E+00	7.49E-01	3.33E+00
Stratospheric ozone depletion	kg CFC11 eq	1.46E-05	7.15E-06	8.46E-06	7.60E-06	7.28E-06	7.92E-06	9.05E-06	8.13E-06	4.76E-06	5.07E-06
Ionizing radiation	kBq Co-60 eq	5.78E-02	9.31E-03	1.61E-02	6.85E-03	1.22E-02	1.88E-02	7.37E-02	6.59E-02	8.65E-03	5.39E-02
Ozone formation, Human health	kg NO_x eq	3.01E-03	1.51E-03	1.78E-03	1.61E-03	1.54E-03	1.66E-03	5.20E-03	4.71E-03	2.91E-03	3.07E-03
Fine particulate matter formation	kg PM2.5 eq	2.46E-03	8.94E-04	1.17E-03	8.90E-04	9.09E-04	1.12E-03	3.17E-03	2.67E-03	1.07E-03	1.94E-03
Ozone formation, Terrestrial ecosystems	kg NO_x eq	3.05E-03	1.53E-03	1.80E-03	1.63E-03	1.55E-03	1.68E-03	5.30E-03	4.76E-03	2.93E-03	3.12E-03
Terrestrial acidification	kg SO_2 eq	1.24E-02	5.16E-03	6.43E-03	5.30E-03	5.24E-03	6.11E-03	1.31E-02	1.13E-02	5.17E-03	7.83E-03
Freshwater eutrophication	kg P eq	6.91E-04	3.52E-04	4.12E-04	3.76E-04	3.58E-04	3.85E-04	9.16E-04	8.41E-04	5.42E-04	5.03E-04
Marine eutrophication	kg N eq	4.68E-03	2.38E-03	2.79E-03	2.54E-03	2.42E-03	2.60E-03	3.24E-03	2.96E-03	1.90E-03	1.78E-03
Terrestrial ecotoxicity	kg 1,4-DCB	5.82E+01	2.96E+01	3.47E+01	3.16E+01	3.01E+01	3.24E+01	7.21E+01	6.62E+01	4.26E+01	3.97E+01
Freshwater ecotoxicity	kg 1,4-DCB	1.03E-01	5.24E-02	6.13E-02	5.59E-02	5.32E-02	5.72E-02	2.52E-01	2.32E-01	1.49E-01	1.39E-01
Marine ecotoxicity	kg 1,4-DCB	1.88E-01	9.59E-02	1.12E-01	1.02E-01	9.74E-02	1.05E-01	7.01E-01	6.43E-01	4.14E-01	3.85E-01
Human carcinogenic toxicity	kg 1,4-DCB	1.68E-03	5.54E-04	8.19E-04	4.26E-04	4.87E-04	8.02E-04	2.44E-03	2.05E-03	8.77E-04	1.51E-03
Human non-carcinogenic toxicity	kg 1,4-DCB	6.02E+00	3.07E+00	3.59E+00	3.28E+00	3.12E+00	3.35E+00	5.26E+00	4.82E+00	3.10E+00	2.89E+00
Land use	m^2a crop eq	6.46E+00	3.29E+00	3.86E+00	3.52E+00	3.34E+00	3.60E+00	1.09E+01	1.00E+01	6.46E+00	6.00E+00
Mineral resource scarcity	kg Cu eq	2.02E-03	9.95E-04	1.18E-03	1.05E-03	1.01E-03	1.10E-03	2.99E-03	2.73E-03	1.70E-03	1.67E-03
Fossil resource scarcity	kg oil eq	9.69E-01	2.17E-01	3.96E-01	1.04E-01	1.72E-01	4.09E-01	1.08E+00	8.52E-01	1.28E-01	8.08E-01
Water consumption	m^3	7.93E-02	3.07E-02	4.35E-02	1.98E-02	2.50E-02	4.25E-02	4.74E-02	3.64E-02	2.78E-03	4.78E-02

Table A 3.6 Environmental impact for fractions obtained from yellow pea and lupine expressed in 1 kg of protein without cultivation included.

Impact category	Unit	Yellow Pea						Lupine			
		CF	MF		DF	HF		CF	MF	DF	HF
		Protein Isolate	Insoluble Protein	Soluble Protein	Fine fraction	Insoluble Protein	Soluble Protein	Protein Isolate	Enriched Protein	Fine fraction	Enriched protein
Global warming	kg CO_2 eq	4.00E+00	8.21E-01	1.95E+00	2.50E-01	9.73E-01	1.79E+00	4.53E+00	3.40E+00	1.86E-01	3.14E+00
Stratospheric ozone depletion	kg CFC11 eq	8.71E-07	1.14E-07	3.08E-07	7.42E-08	2.22E-07	2.94E-07	1.22E-06	9.16E-07	5.53E-08	7.80E-07
Ionizing radiation	kBq Co-60 eq	6.62E-02	1.16E-02	2.39E-02	8.83E-03	2.49E-02	2.52E-02	7.53E-02	6.75E-02	6.58E-03	5.68E-02
Ozone formation, Human health	kg NO_x eq	5.91E-05	6.57E-06	1.55E-05	1.87E-06	7.51E-06	1.41E-05	3.29E-04	2.23E-04	1.40E-06	4.31E-04
Fine particulate matter formation	kg PM2.5 eq	1.13E-03	1.64E-04	4.28E-04	7.49E-05	2.51E-04	3.96E-04	1.62E-03	1.22E-03	5.58E-05	1.12E-03
Ozone formation, Terrestrial ecosystems	kg NO_x eq	7.93E-05	1.06E-05	2.49E-05	3.02E-06	1.21E-05	2.28E-05	3.99E-04	2.45E-04	2.25E-06	4.59E-04
Terrestrial acidification	kg SO_2 eq	3.63E-03	5.22E-04	1.37E-03	2.38E-04	7.96E-04	1.27E-03	5.25E-03	3.95E-03	1.77E-04	3.60E-03
Freshwater eutrophication	kg P eq	2.55E-07	2.91E-08	7.64E-08	1.96E-08	5.81E-08	7.37E-08	4.50E-07	3.58E-07	1.46E-08	3.81E-07
Marine eutrophication	kg N eq	1.93E-05	9.54E-07	5.57E-06	3.99E-07	1.70E-06	4.46E-06	3.61E-05	2.31E-05	2.98E-07	1.84E-05
Terrestrial ecotoxicity	kg 1,4-DCB	1.65E-01	2.23E-02	5.12E-02	1.60E-02	4.62E-02	5.17E-02	2.59E-01	2.17E-01	1.19E-02	2.12E-01
Freshwater ecotoxicity	kg 1,4-DCB	6.95E-05	1.27E-05	2.76E-05	7.06E-06	2.24E-05	2.72E-05	8.84E-05	7.35E-05	5.26E-06	7.08E-05
Marine ecotoxicity	kg 1,4-DCB	3.50E-04	6.40E-05	1.43E-04	3.04E-05	9.97E-05	1.38E-04	5.12E-04	4.24E-04	2.26E-05	4.98E-04
Human carcinogenic toxicity	kg 1,4-DCB	1.18E-03	3.11E-04	6.99E-04	4.42E-05	2.67E-04	6.21E-04	1.14E-03	8.31E-04	3.29E-05	8.19E-04
Human non-carcinogenic toxicity	kg 1,4-DCB	8.47E-03	1.45E-03	3.26E-03	8.13E-04	2.51E-03	3.20E-03	1.24E-02	9.45E-03	6.06E-04	9.40E-03
Land use	m^2a crop eq	4.95E-04	1.45E-04	3.33E-04	0.00E+00	7.60E-05	2.85E-04	5.60E-04	4.05E-04	0.00E+00	5.95E-04
Mineral resource scarcity	kg Cu eq	1.21E-04	2.46E-05	5.41E-05	1.03E-05	3.64E-05	5.14E-05	1.47E-04	1.17E-04	7.64E-06	1.16E-04
Fossil resource scarcity	kg oil eq	1.05E+00	2.58E-01	5.82E-01	5.91E-02	2.62E-01	5.29E-01	1.04E+00	8.01E-01	4.40E-02	8.19E-01
Water consumption	m^3	5.48E-02	2.22E-02	4.13E-02	4.50E-04	1.68E-02	3.60E-02	4.95E-02	3.75E-02	3.35E-04	5.22E-02

Table A 3.7 Environmental impact for fractions obtained from yellow pea and lupine expressed in 1 kg of protein with cultivation included.

Impact category	Unit	Yellow Pea						Lupine			
		CF	MF		DF	HF		CF	MF	DF	HF
		Protein Isolate	Insoluble Protein	Soluble Protein	Fine fraction	Insoluble Protein	Soluble Protein	Protein Isolate	Enriched Protein	Fine fraction	Enriched protein
Global warming	kg CO_2 eq	5.33E+00	1.78E+00	3.12E+00	1.58E+00	2.42E+00	2.72E+00	5.78E+00	4.55E+00	1.30E+00	3.83E+00
Stratospheric ozone depletion	kg CFC11 eq	1.85E-05	1.29E-05	1.59E-05	1.77E-05	1.95E-05	1.28E-05	1.04E-05	9.40E-06	8.27E-06	5.83E-06
Ionizing radiation	kBq Co-60 eq	7.33E-02	1.68E-02	3.02E-02	1.60E-02	3.27E-02	3.02E-02	8.48E-02	7.62E-02	1.50E-02	6.19E-02
Ozone formation, Human health	kg NO_x eq	3.82E-03	2.72E-03	3.33E-03	3.76E-03	4.11E-03	2.67E-03	5.98E-03	5.44E-03	5.05E-03	3.54E-03
Fine particulate matter formation	kg PM2.5 eq	3.12E-03	1.61E-03	2.19E-03	2.07E-03	2.43E-03	1.81E-03	3.64E-03	3.09E-03	1.87E-03	2.23E-03
Ozone formation, Terrestrial ecosystems	kg NO_x eq	3.87E-03	2.75E-03	3.37E-03	3.80E-03	4.15E-03	2.70E-03	6.09E-03	5.50E-03	5.09E-03	3.58E-03
Terrestrial acidification	kg SO_2 eq	1.57E-02	9.28E-03	1.21E-02	1.24E-02	1.40E-02	9.82E-03	1.51E-02	1.30E-02	8.98E-03	9.01E-03
Freshwater eutrophication	kg P eq	8.76E-04	6.33E-04	7.74E-04	8.77E-04	9.56E-04	6.19E-04	1.05E-03	9.73E-04	9.42E-04	5.79E-04
Marine eutrophication	kg N eq	5.94E-03	4.28E-03	5.23E-03	5.92E-03	6.46E-03	4.19E-03	3.72E-03	3.43E-03	3.30E-03	2.04E-03
Terrestrial ecotoxicity	kg 1,4-DCB	7.39E+01	5.33E+01	6.51E+01	7.38E+01	8.05E+01	5.21E+01	8.30E+01	7.66E+01	7.39E+01	4.56E+01
Freshwater ecotoxicity	kg 1,4-DCB	1.30E-01	9.42E-02	1.15E-01	1.30E-01	1.42E-01	9.21E-02	2.90E-01	2.68E-01	2.59E-01	1.59E-01
Marine ecotoxicity	kg 1,4-DCB	2.39E-01	1.72E-01	2.11E-01	2.39E-01	2.60E-01	1.69E-01	8.06E-01	7.44E-01	7.20E-01	4.43E-01
Human carcinogenic toxicity	kg 1,4-DCB	2.13E-03	9.96E-04	1.54E-03	9.93E-04	1.30E-03	1.29E-03	2.81E-03	2.37E-03	1.52E-03	1.73E-03
Human non-carcinogenic toxicity	kg 1,4-DCB	7.64E+00	5.52E+00	6.74E+00	7.64E+00	8.33E+00	5.39E+00	6.05E+00	5.58E+00	5.39E+00	3.32E+00
Land use	m^2a crop eq	8.19E+00	5.92E+00	7.23E+00	8.20E+00	8.94E+00	5.79E+00	1.26E+01	1.16E+01	1.12E+01	6.90E+00
Mineral resource scarcity	kg Cu eq	2.56E-03	1.79E-03	2.21E-03	2.45E-03	2.70E-03	1.78E-03	3.44E-03	3.16E-03	2.95E-03	1.93E-03
Fossil resource scarcity	kg oil eq	1.23E+00	3.90E-01	7.43E-01	2.41E-01	4.61E-01	6.58E-01	1.24E+00	9.85E-01	2.23E-01	9.29E-01
Water consumption	m^3	1.01E-01	5.53E-02	8.17E-02	4.63E-02	6.68E-02	6.83E-02	5.46E-02	4.21E-02	4.84E-03	5.49E-02

Chapter 4

Concept

Functionality-driven food product formulation – An illustration on selecting sustainable ingredients building viscosity

---

A. Lie-Piang, A.C. Möller, N. Köllmann, A. Garre, R.M. Boom, A. van der Padt

Food Research International (2022)

Abstract

Currently, food industries typically favour formulation of food products using highly refined techno-functional ingredients of high purity. However, there is a growing interest in less pure techno-functional ingredients with a lower degree of refining as they deliver the same functional properties with reduced environmental impact. We propose that instead of selecting formulations based on purity, they should be selected based on their techno-functional properties. This article illustrates that the shift in perspective may increase the sustainability of food production. The functionality-driven product formulation is explored through a case study in which yellow pea ingredients are selected to increase the viscosity of a salad dressing. The relation between the ingredients (in terms of composition; protein, starch fibre, and a residual fraction) and the final viscosity was quantified and validated using multiple linear regression. The model described the observations well: the final viscosity is mostly dominated by the starch content; protein content has only a marginal impact; and dietary fibre contributes to viscosity with an antagonistic effect with starch. Based on the multiple linear regression model and further formulation optimisation, we identified various combinations of ingredients (with either a high or low degree of refining) that would result in the target final viscosity. An evaluation of the global warming potential of all blends showed that the desired viscosity could be achieved using only isolates, as well as by using only mildly refined fractions. The latter is associated with a global warming potential that is 80% lower than the one based on isolates. This case study demonstrates the proof of concept for this approach, showing it can aid in identifying alternative product formulations with similar techno-functional properties but with a higher sustainability.

4.1 Introduction

Food products are commonly assembled from ingredients with high purity such as protein or starch isolates. These ingredients are known to deliver a certain techno-functional property with reliable quality, yet their production is also connected to the generation of low-quality side streams as well as high resource use. Ingredients with a lower degree of refining produced by for example dry or mild aqueous fractionation are, besides their clean label, promising alternatives to the pure ingredients with good techno-functional properties such as thickening[65], gelling[99], and foaming capacity[136]. Moreover, the global warming potential of mild aqueous and dry fractionation is respectively ~45 and ~90% lower compared to the conventional full refining.[9] However, less refined ingredients are less pure compared to highly refined ingredients. As ingredients are normally selected based on composition, e.g., starch for thickening and proteins for emulsions, the implementation of the milder multi-component ingredients into formulated foods is not straightforward. Jonkman et al.[18] showed that the implementation of multi-component ingredients in the food industry can be facilitated by blending different ingredients to fulfil the requirements of the food products. It was proposed that by-products can also be utilized if ingredients were selected based on their techno-functional properties instead.

To accomplish a selection of blends based on techno-functional properties, one should be able to derive this functionality based on the characteristics of the ingredients. In other words, the contribution of the different ingredients to the techno-functional properties should be quantified. The main challenge is to pinpoint the contribution of single ingredients in the blends to the functional properties as they vary a lot in composition. In addition, the interactions between the components and the differences in functionality between ingredients with a high or a low degree of refining have to be considered. Models such as the Krieger-Dougherty equation in combination with other models[145] and relationships such as the power law[65] can be used to describe the effective viscosity or the storage modulus[146] in food systems. However, these models cannot distinguish between interactions of multiple components and are, hence, less suitable to describe techno-functional properties of ingredients blends with complex compositions. Therefore, we assert that these types of models are not applicable to select blends of complex ingredients.

The use of mathematical models has been employed for estimating microbial risk in food products[147] and for the formulation of low sugar or fat products in product development[20,148]. Monnet et al.[149] applied multi-objective optimisation using regression models to predict the effect of variations in flour quality on the processing variables of baking cakes. These studies all make use of empirical models to quantify relationships between the relevant variables, either safety parameters, microbial indicators, or process attributes. Other studies have used sustainability as a factor in multi-objective decision models, such as the manufacturing industry[150] and biomass supply chain[151]. A combination of the aforementioned approaches could result in food product formulation that is optimised in terms of both sustainability and techno-functionality. Therefore, we aim here to assess the potential of regression for quantifying the contribution of the ingredients to techno-functional properties, to facilitate the functionality-driven selection of sustainable ingredients that are less refined. Initially, blends of commercial isolates (CI) and mildly refined yellow pea ingredients (MRF) were created to compare the techno-functionality of the two types of ingredients and investigate their contribution to the final viscosity. Then, a multiple linear regression model was fitted to the data, allowing us to predict the final viscosity based on composition. The practical relevance is illustrated with a case study, using the model to select a combination of ingredients that results in the desired techno-functional properties, with the lowest environmental impact.

4.2 Materials & methods

Pre-dried yellow peas (*Pisum sativum L.*), were purchased from Alimex (The Netherlands). Commercial pea protein isolate (Nutralys F85G) and pea fibre isolate (PEA FIBRE I 50 M) were purchased from Roquette Frères S.A (St. Louis, USA). Pea starch isolate was purchased from Emsland Stärke GmbH (Germany). For all experiments deionized water produced with a Milli-Q purification system was used (Merck Millipore, Burlington, USA).

4.2.1 Production of mildly refined ingredients

To obtain non-dehulled pea flour, non-dehulled peas were pre-milled into grits using a pin mill (LV 15 M, Condux-Werk, Germany) at room temperature. The grits were subsequently milled into fine flour using a ZPS50 impact mill (Hosokawa-Alpine,

Germany), with an impact mill speed of 8000 rpm, the air flow at 52 m^3/h, the classifier wheel speed at 4000 rpm, and a feed rate of 2 rpm (method adopted[49]). A thermometer inside the mill was used to control the temperature between 16 and 34 ˚C.

For the dry fractionated ingredients, the peas were initially dehulled with a Satake TM05 pearling machine (Satake Corporation, Japan). Subsequently, they were pre-milled and milled according to the non-dehulled peas. Yellow pea ingredients enriched in protein and starch were obtained by air classification, according to the method of Pelgrom et al. [52] An ATP50 classifier (Hosokawa-Alpine, Germany) was used, with a fixed air flow at 52 m^3/h, the classifier wheel speed at 5000 rpm, and the feed rate at 20 rpm.

4.2.2 Compositional analysis

The protein content of the MRFs was determined with Dumas analysis (Thermo Fischer Scientific, Massachusetts, USA) and that of the CIs using Kjedahl method. A comparison between the protein contents obtained for the isolates by both methods showed no difference. For both methods a nitrogen conversion factor of 5.52 for pea protein was used.[137] The total starch content was determined with a Total Starch Amyloglucosidase-α-Amylase Assay Kit, AOAC Method 996.11 (Megazyme International Ireland Ltd, Bray, Ireland) and the total dietary fibre content in the MRFs was measured using a Total Dietary Fibre Assay Kit, AOAC Method 991.43/AACC Method 32–07.01 (Megazyme International Ireland Ltd, Bray, Ireland) and the fibre in the CIs was determined using the AOAC method 985.29. These two AOAC methods are very similar; 985.29 measures the total dietary fibre and 991.43 measures total dietary fibre separated into insoluble and soluble fibre.[152] Dry matter contents were determined by drying with an infrared moisture analyser (MA35, Sartorius, Göttingen, Germany) at 120 °C. The residual components were determined as the difference between the dry matter and the protein, starch, and fibre content and contain among others oil, ash, salts, and sugars.

4.2.3 Formulation & sample preparation

Protein, starch, and fibre isolates and the dry fractionated pea flour and protein- and starch-rich ingredients were formulated in different ways:

- single isolates and mixtures of only isolates,
- single mildly refined ingredients and mixtures thereof, as well as
- mixtures of isolates and mildly refined ingredients.

The formulations of the samples were chosen such that the individual ingredients cover a range of dry matter concentrations as well as compositions. The mildly refined ingredients were also mixed such that their compositions were similar to those of the mixes from the isolates. The complete formulation table is provided in the Appendix (Table A 4.1). Next, the dataset was divided into a training set (n = 95) used for parameter estimation and an independent validation set (n = 15). Figure 4.1 illustrates the compositions in fibre, starch, and protein of the samples in the training and validation sets, the size of the bubbles represents the relative concentration. The residual components (i.e. oil, sugars, and ash) are not depicted in this Figure but are considered in the modelling.

To prepare the samples for the viscosity measurements, the dry ingredients were initially pre-mixed. The dry mixes were then added to water in a concentration range of approximately 2-22 wt% on a dry base. The samples were continuously hydrated using a magnetic stirrer for 5 – 10 min (depending on the consistency of the dispersion) at room temperature before the viscosity measurements.

4.2.4 Viscosity measurement

The viscosity of the dispersion was determined with an Anton Paar Rheometer MCR301 and a starch cell C-ETD160/ST (Anton Paar, Austria). The dispersions were stirred at 600 rpm for 10 s at 50 °C for to reassure a thorough dispersion prior to starting the measurement. The viscosity measurements were performed at a constant rotational speed of 160 rpm, and a temperature–time profile was used for all samples (Figure A 4.1, Appendix). The temperature time profile is a typical pasting profile for Rapid-Visco analysers which was adapted to be used in a

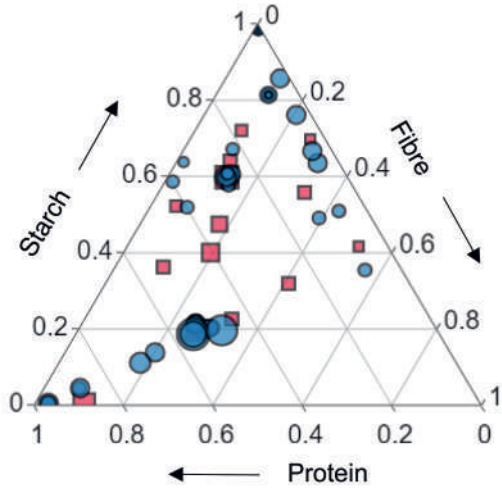

Figure 4.1 Overview of samples that were measured in the training set (●) and in the validation set (■). The ternary diagram gives a distribution of the starch, protein, and fibre content, please be aware there is also a rest ingredient in these samples. The size of the symbols represents the concentration of the samples ranging from approximately 2 -20 wt%.

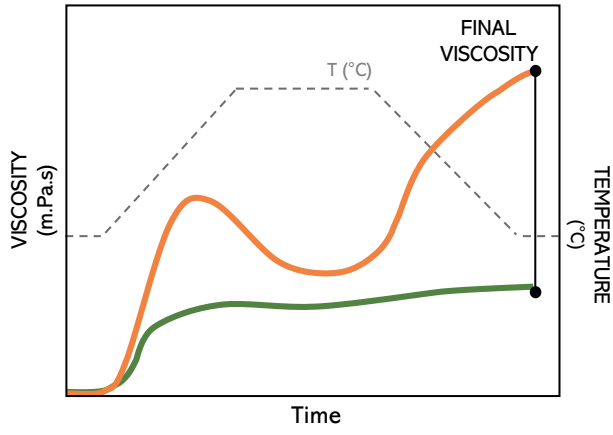

Figure 4.2 Illustration for the determination of the final viscosity of two typical viscosity profiles. The samples are initially heated to 95 °C, after which they are cooled again to 50 °C and the final viscosity is determined.

rheometer.[153] The final viscosity was calculated with the Anton Paar Rheocompass software (Anton Paar, Austria) and is determined at the end of the cooling phase. As the viscosity profiles vary among the ingredients, the point at which the final viscosity is determine is illustrated in Figure 4.2. All viscosity determinations were performed at least in duplicate.

4.2.5 Modelling methods

4.2.5.1 Describing final viscosity as dependent variable

The relationship between the final viscosity of the product and its properties was described using adapted multiple linear regression. A multiple linear regression model can be described according to Equation (4-1), where η is the dependent variable, and X_i are the values of the explanatory variables. The individual effect of each explanatory variable is quantified by the slope coefficients, b_i; with larger values indicating a stronger effect. The combined effect of variables X_i and X_j is defined by coefficients c_{ij}; with $c_{ij} > 0$ indicating a synergistic effect and $c_{ij} < 0$ representing an antagonistic effect. Finally, η_0 is the intercept of the model and in our case would represent the viscosity when all explanatory variables are zero: $X_i = 0 \; \forall \; i$.

$$\eta = \eta_0 + \sum_i b_i X_i + \sum_{i \neq j} c_{ij} X_i X_j \qquad (4\text{-}1)$$

On a first step, an exploratory analysis was performed to evaluate the data transformation more suitable for the hypotheses of a linear model. Based on graphical representations and basic statistical analyses (not shown), we concluded that a log-transformation on the final viscosity resulted in residuals with constant variance and without significant deviations from normality. Therefore, the log-transformed viscosity was used as dependent for the regression as is depicted in Equation (4-2).

$$ln\, \eta = \eta_0 + \sum_i b_i X_i + \sum_{i \neq j} c_{ij} X_i X_j \qquad (4\text{-}2)$$

Three families of regression models based on Equation (4-2) were developed within this study. The first family described the (log-transformed) final viscosity of the product based on the dry matter content of the product. Therefore, this model has a

single explanatory variable: the dry matter content in weight %. The second family of models describes the (log-transformed) final viscosity based on the content of each ingredient in weight %. Therefore, the explanatory variables of this model describe the flour, protein isolate, starch isolate, fibre isolate, starch-rich ingredient, and protein-rich ingredient content. Potential interactions within this model were excluded to prevent overfitting. Finally, the third family of models describes the (log-transformed) final viscosity based on the model components. This model has as explanatory variables the protein, starch, and fibre content (in dry matter weight %). Moreover, an additional explanatory variable "rest" was included in this model representing the remaining dry weight.

The most suitable model structure was determined by model selection, using an approach adapted from González-Tejedor et al.[147] When building empirical models, it is desirable to use models able to describe the data with a minimum number of model parameters; i.e. the most parsimonious model.[154] For that reason, every possible variation of Equation (4-2) with one or more coefficient fixed to zero was fitted to the model. Then, the most parsimonious model was selected based on the Akaike Information Criterion (AIC), following Gonzalez-Tejedor et al.[147] All these operations were done using R version 3.6.3. The R code is available in the corresponding git library available at https://git.wur.nl/anouk.lie-piang/functionality-oriented-selection-of-ingredients.git. The model was subsequently transformed to come closer to a physical model by forcing the intercept through 1 mPa.s, i.e. the viscosity of water, thus $\ln \eta_0 = 0$. The adjusted R^2 was subsequently calculated using equation (4-3).

$$\text{Adjusted } R^2 = 1 - \frac{SSR/(n-p-1)}{SST/(n-1)} \tag{4-3}$$

In which SSR is the sum of squares related to the residuals, the SST is the total sum of squares, n is the total number of samples, and p is the number of explanatory variables (without the intercept).

The models were validated using independent data points not used for model fitting. Namely, 95 points were used for training the models and 15 for model validation. More details about the quality of the regression models were revealed by calculating

the R^2 using Equation (4-4), as opposed to the adjusted R^2 for the model selection, as the purpose here was rather to get an indication of the performance of the initial model on a new dataset instead of penalizing variables.

$$R^2 = 1 - \frac{SSR}{SST} \tag{4-4}$$

In addition, the residuals are calculated using Equation (4-5) to pinpoint the quality of the prediction per sample.

$$Residuals_i = (y_i - \hat{y}_i)^2 \tag{4-5}$$

The accuracy (i.e. variance) of the validation set was compared with the training set model by using the residual standard error (RSE) calculated using Equation 4-6.

$$RSE = \sqrt{\frac{SSE}{n-p-1}} \tag{4-6}$$

4.2.5.2 Selection of sustainable ingredients

The model with the best description of the data was used to select the most sustainable blends of ingredients based on the final viscosity they can deliver. First, the regression model description was used to generate all possible combinations of starch, fibre, protein, and rest that would result in a viscosity of 1,500 +/- 2.5 % mPa.s in Python version 3.8.3 using the module python-constraint 1.4.0. This module solves constraint satisfaction problems over finite domains, and takes a variable problem, here the required final viscosity, that can be solved using additional variables (the compositions). These variables are subsequently constrained, for example by upper or lower limits. The solutions is a list of combinations of all variables (in this case compositions), that adhere to the constraints and do result in the required final viscosity. As these compositions belong to a continuous formulation window, this would generate an infinite amount of numbers. Therefore, steps of 0.1 wt% in dry matter were taken (e.g. 1.1, 1.2, 1.3 wt% of protein), to scan the allowed formulation window. The following constraints were applied:

1) the compositions generated should result in an output of 1,500 mPa.s +/- 2.5 %,

2) the maximum dry matter allowed is 22 wt% as this was the upper range in the training set, and

3) upper and lower boundaries were set for the dry matter concentration of protein, starch, fibre and rest as they are represented in the training set.

The list of allowable compositions obtained from the constraint model were subsequently subjected individually to linear programming using the python PuLP module. The objective of the model is to minimise the quantity of a blend of fractions, subjected to the constraints of the composition generated in the constraint model. Now, the total sum of ingredients in a blend that matches the combination of components (achieving a viscosity of 1,500 mPa.s) was minimised. Lastly, the global warming potential (GWP) per ton (1000 kg) product (the blends of ingredients) are calculated using data from a previous life cycle assessment summarized in Table 4.1.[9] The total GWP for ton blend containing a mass ingredient of ingredient f was calculated by using Eq. (4-7).

$$GWP = \frac{\sum_f X_f GWP_f}{\sum_f X_f} \cdot 1000 \qquad (4\text{-}7)$$

Table 4.1 Global warming potential (GWP) of yellow pea ingredients using system boundaries of cultivation and processing.[9supplementary information]

Yellow Pea	Protein isolate	Starch isolate	Fibre isolate	Starch-rich ingredient	Protein-rich ingredient	Flour
GWP CO_2-eq/kg ingredient	4.2	1.0	0.86	0.68	0.68	0.67

4.3 Results & discussion

4.3.1 Observed final viscosity of yellow pea ingredients

To allow for a functionality-driven selection of ingredients, the relation between the ingredients and their techno-functionality, the final viscosity in this case, needs to be investigated first. The final viscosity could be described by the total amount of dry matter, the quantity of the ingredients, or by the components (e.g. protein, starch, fibre, and the rest); all of which are explored in this study. The viscosity was determined of both commercial and mildly refined enriched yellow pea ingredients, by measuring the final viscosity of the ingredient dispersions in water. A typical

viscosity profile of each single ingredient at a concentration of 10 wt% is depicted in Figure 4.2. The ingredients were divided into three categories: commercial isolates, both single and mixtures (CI), mildly refined ingredients, single or mixtures (MRF), and mixtures of both isolates and MRFs.

Table 4.2 lists the compositions of the MRFs and the CIs. The MRFs are of course less pure compared to the isolates. The compositional analysis allows us to allocate the final viscosity increase to each ingredient, but also to the contributions of the individual components.

The contributions of the single ingredients to the viscosity are depicted in Figure 4.3A. For CIs and MRFs, the final viscosity increases as a function of dry matter, which follows an initial exponential trend that levels off at higher dry matter contents. However, the magnitude of the increase varies between ingredient; not all ingredients contribute to the viscosity to the same extend. The viscosity of starch isolate increases faster than the viscosity of protein isolate. Only one duplicate of fibre isolate was measured, and therefore no trend of the contribution is depicted. Nevertheless, it is clear that the viscosity increases with dry matter, with the least contributor being protein isolate, then fibre isolate, and starch isolate being the strongest contributor. The viscosity contributions of the MRFs are alike. The viscosity profiles are depicted for the starch-rich ingredient, flour, and protein-rich ingredient. The viscosity contribution of the ingredients is increasing from the protein-rich ingredient, to flour, to starch-rich ingredient, respectively. As with the CIs, the starch-rich ingredient has a greater thickening capacity than the protein-rich ingredients. Therefore, in general, an increase in starch and decrease in protein results in an increase the final viscosity. This is expected and in line with previous results, in which substitution of starch by protein resulted in an overall decrease in the apparent viscosity profile.[155] Flour is the combination of both, so it is natural that this ingredient lays in between the mildly fractionated variants of it.

When comparing the final viscosities of CIs versus MRFs, one can observe that the viscosity contribution of the starch isolate is approximately double the contribution of the starch-rich ingredient at the same dry matter content. In addition, the contribution of the dry fractionated protein-rich ingredient to the final viscosity

Table 4.2 Composition of commercial isolates and mildly refined enriched ingredients, with a 95% confidence interval.

	Dry matter (g/100g sample)	Total Starch (g/100g dry matter)	Protein (g/100g dry matter)	Total Dietary Fibre (g/100g dry matter)	Residual components[a] (g/100g dry matter)
Starch-rich ingredient	90.2 ± 0.4	62.6 ± 1.9	15.3 ± 0.6	9.0	13.2 ± 2.1
Protein-rich ingredient	91.0 ± 1.5	13.8 ± 4.9	40.7 ± 1.8	19.4	26.0 ± 5.8
Non-dehulled Flour	91.1 ± 0.1	47.9 ± 7.1	21.6 ± 0.5	10.6	19.9 ± 7.2
Fibre isolate	95.2 ± 0.5	33.8 ± 0.5	7.83 ± 0.5	53.6 ± 4.6	4.84 ± 1.17
Protein isolate	93.7 ± 0.5	0.39 ± 0.3	74.1 ± 1.8	2.24 ± 1.2	23.3 ± 1.9
Starch isolate	88.4 ± 0.3	91.4 ± 2.9	< 0.5	0.79 ± 0.7	< 7.27

[a] Calculated by difference total dry matter starch content, protein content with conversion factor 5.52 and fibre content.

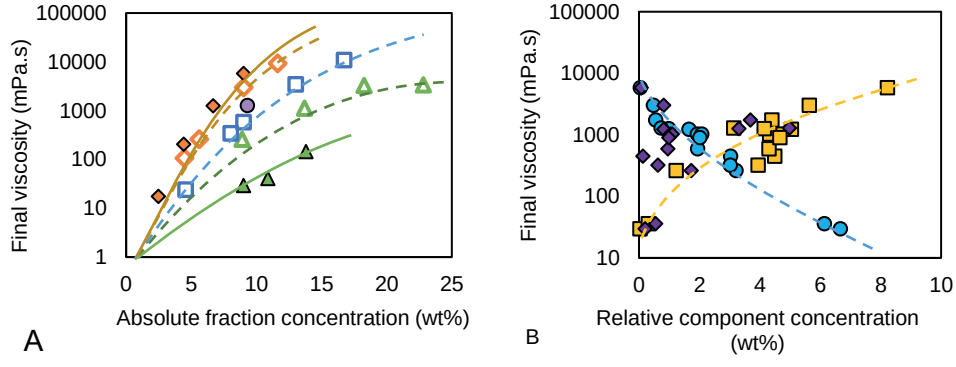

Figure 4.3 A) Final viscosity of starch (♦), protein (▲), fibre (●) isolates and the milder refined starch-rich (◊) and protein-rich (△) ingredients and flour (□) as a function of concentration (wt%) of the fractions. b) 15 dispersions of CIs, MRFs, and mixtures with a constant concentration of 8.8-9.2 wt% split into the relative protein (●), starch (■) and fibre (♦) content. Lines are to guide the eye.

compared to the protein isolate is an order of magnitude larger at similar dry matter contents. Both can be attributed to the amount of starch in the respective ingredient. Starch isolate has a larger starch content than the dry fractionated starch-rich ingredient, which also contains other components such as protein and dietary fibre. Similarly, protein isolate is purer than the dry fractionated protein-rich ingredient, and contains almost no starch, while the protein-rich ingredient contains approximately 14% starch on a dry base (Table 4.2). In addition, the protein-rich ingredient also contains a considerable amount of fibre, which could increase the viscosity as well. To gain more insight in the contribution of CIs and MRFs to the final viscosity, all ingredients (CI-CI, MRF-MRF, and CI-MRF) were mixed together (Figure A 4.2, Appendix). For all ingredients and mixtures, we see a similar trend for increasing viscosity with increasing dry matter content.

To assess the component contribution to the final viscosity, 15 dispersions of CIs, MRFs, and mixtures with similar concentrations (8.8–9.2 wt%) were plotted with their composition against the final viscosity (Figure 4.3B). It was assumed that within this dry matter range, the influence of the dry matter content on the final viscosity could be neglected. The 15 samples are now split into the relative protein, starch, and fibre content. The Figure implies that in the samples used, starch and protein substitute each other. Hence, more protein means less starch and vice versa, explaining the decrease in viscosity with increasing protein content. With regards to the fibre contribution to the viscosity, two trends can be seen in this dry matter range. Higher viscosities were obtained with both higher and lower fibre content. Three data points depict dispersions that contain more than 3 wt% fibre and contribute to the same viscosities as samples containing only around 1 wt% fibre. The mixtures containing more than 3 wt% fibre, all contain fibre isolate, while the other mixtures do not contain fibre isolate; instead in those mixtures the fibre is inside MRFs. Hence, a difference in contribution to viscosity seems to exist between the fibre originating from fibre isolate and that in the MRFs. As the fibre isolate as well as the dry fractionated fractions are dehulled, it is not expected that the type of fibre is different. Effects of processing history could induce a difference between the functionality of fibre, such as the drying step in fibre isolation, which is absent the production of dry fractionated ingredients. A study that investigated the effect of

heating on dietary fibre in wheat and barley showed that heating may cause a redistribution of the soluble and insoluble components in fibre and subsequently increase the viscosity.[156] In this case that would mean that a fibre isolate has a larger effect on viscosity than fibre in mildly refined components. However, it could also be that at low (1 wt%) and higher (3 wt%) fibre concentrations, other interactions take place with the starch, protein, or residual fraction. This highlights the complexity of pinpointing the contribution of components to the final viscosity based only on mechanistic considerations, and demonstrates why a data-driven approach is preferable for our purposes. Hence, to employ functionality-driven ingredient selection, mathematical models can provide further insight on relevant mechanisms, in which a distinction between the processing history of ingredients can be discerned.

4.3.2 Modelling the final viscosity of yellow pea ingredients

4.3.2.1 Describing viscosity using regression models

So far it was shown that there is a clear relation between the logarithm of the final viscosity and the dry matter content. Especially samples containing starch and fibre contribute significantly to the viscosity, whereas substitution with proteins results in a reduced final viscosity. Nevertheless, these conclusions are rather qualitative and can hardly be used to predict the final viscosity of ingredient blends. Therefore, the data were fitted using multiple linear regression to quantify and predict the final viscosity. Three explanatory variable categories could be established: the dry matter level, the individual ingredients, and the components within the ingredients. The components in the ingredients can be divided into starch, protein, fibre, and the rest. The latter contains all that is not included in the main components, such as sugars, oil, and ash. The best models for each category were selected based on the AIC and are summarized in Table 4.3. The fitting of each model is also depicted in Figure A 4.3, including the adjusted R^2.

The first model, with the dry matter percentage of the yellow pea ingredients as the explanatory variable, had a very limited descriptive value, if any. The second model with the ingredients as descriptive variables could describe the data much better. The values of the regression coefficients do indicate that the starch and fibre isolates

and the starch-rich ingredients contribute most to the final viscosity, which is in accordance with our earlier observations (Figure 4.3). The model based on the components in the ingredients was slightly more descriptive with an adjusted R^2 of 0.93. Notably, the interpretation of the regression coefficients in the component model as a single value should be used with caution, as these are influenced by the interactions as well. Nevertheless, it is clear that fibre and starch contribute the most to the final viscosity, which coincides with the earlier findings on the observational viscosity data. The component model indicates antagonistic interaction between starch and fibre and between starch, fibre, and protein; the presence of two or three components together decreases the final viscosity. These interactions cannot be directly translated into quantitative, physical interactions, as interactions terms are mostly used to fit data better. As the purpose of this study is solely to predict the functional properties to select ingredients rather than understanding and characterising the physical interactions, this is acceptable in the context of this paper. The presence of interactions does however indicate that the viscosity is a complex phenomenon that cannot be fully described by the sole effect of single components. These interactions are investigated in a follow-up study using an adapted polymer blending law.[153] The fact that interactions are required to obtain a good description, demonstrates that isolated mechanistic models like the Krieger Dougherty model would also not be suitable in this specific case to describe the final viscosity without adding extra empirically determined interaction parameters. Although the contribution of each component was expected (e.g. the high contribution of starch), the quantification now allows for the incorporation of the techno-functional properties in optimisation models to select the functional ingredients, as is shown in the last part of this study.

The intercept of the multiple regression model represents the viscosity of the product when the concentration of every component is zero; i.e. when it is just water. Even though the current intercept of the model is 1.22 ± 0.23 ln mPa.s, which is not too far from 1 mPa.s (the viscosity of water at room temperature[157]), it is still different from the physical value. To make the model consistent with known extremes, the intercept of the component model including all data was forced through 0 ln mPa.s (resulting in 1 mPa.s) (Table 4.3) The model parameters did not change a lot

Table 4.3 Regression coefficients of the log transformed linear models describing the viscosity by dry matter, ingredients, and components in dry matter percentages. The component model was also fitted with an intercept fixed at 1 mPa.s ($\ln \eta_0 = 0.$). Significance levels are indicated *with '***', '**', '*', '.', and ' ', are 0, 0.001, 0.02, 0.05, and 0.1 respectively.*

Dry matter model			Ingredient model			Component model			Component model intercept	
Variables	Estimate	SD	Variables	Estimate	SD	Variables	Estimate	SD	Estimate	SD
Intercept	4.00 ***	0.39	Intercept	2.57 ***	0.19	Intercept	1.22 ***	0.23	0.00	0.00
Dry matter	0.24 ***	0.03	Starch-rich ingredient	0.54 ***	0.03	Protein	0.13 **	0.04	0.27 ***	0.04
			Protein-rich ingredient	0.32 ***	0.02	Starch	0.87 ***	0.05	1.09 ***	0.03
			Flour	0.40 ***	0.02	Fibre	1.36 ***	0.12	1.78 ***	0.10
			Protein isolate	0.14 ***	0.02	Rest	0.74 ***	0.11	0.80 ***	0.13
			Starch isolate	0.60 ***	0.03	Protein:Starch	-0.06 ***	0.02	-0.09 ***	0.02
			Fibre isolate	0.47 ***	0.02	Protein:Fibre	-0.12 ***	0.02	-0.18 ***	0.03
						Starch:Fibre	-0.18 ***	0.02	-0.27 ***	0.02
						Protein:Starch:Fibre	0.03 ***	0.01	-0.04 ***	0.01
Adjusted R^2	0.34		0.90			0.93			0.91	

relatively to each other, and still describes the viscosity very well (adjusted R2 = 0.91). That the model can still describe the relationship between final viscosity and components, is a strong indication that the initial model was already close to representing the underlying mechanisms of viscosity. Nevertheless, the model with the intercept through 1 mPa.s will be further discussed and used in this study to predict the final viscosity of blends of yellow pea ingredients as it contains the physical boundary.

4.3.2.1.1 Consideration of processing history ingredients

From the previous section, we conclude that the final viscosity can be described based on the yellow pea components from both CI and MRF. However, the proteins in CIs are precipitated and spray dried and therefore considered non-native whereas the MRFs are not; hence, the proteins in the latter are still in their native state. For example, the protein nativity was found to play a role in the emulsifying[98] and gelling capacity[99] of yellow pea ingredients. Therefore, the residuals of the component model (with intercept fixed at 1 mPa.s) were separated according to their origin (Figure A 4.4). They did not show any trend for either CI or MRF samples. As an additional test, a linear regression model was fitted with the protein split into native (originating from MRFs) and non-native protein (originating from CIs). This resulted in a model description close to the original approach, yet with higher variance (not shown). Based on this we conclude that for our empirical design, there are no differences between the final viscosity of native versus non-native protein. This agrees with the observation reported in Figure A 4.2 (Appendix), in which no differences between CIs and MRFs were observed. There may be several explanations for this. The first one is the relatively low protein content of MRFs of approximately 40% of the total dry matter, as opposed to approximately 74% of total dry matter in isolates, which might be too low to have any difference caused by the different degree in the nativity. However, one should also consider that during the pasting procedure, all our mixtures were heated to 95 °C for five minutes, which is above the denaturation temperature of around 83 °C.[99] Therefore, most proteins that were still native, will have become denatured as well. Similar analyses were performed for the differences between fibre from MRF and CIs and resulted in the same conclusions as for protein (data not shown). Therefore, it is suggested that the

difference in fibre from CIs and MRFs mainly originates from the interactions that take place at different concentrations.

4.3.2.2 Validation of multiple linear regression model

We showed that the final viscosity of both CI and MRF ingredients of yellow pea can be described well by the components starch, fibre, protein, and rest, without distinguishing between the processing history. The model was subsequently validated using independent data points not used for model fitting. The final viscosity was measured for a validation set of 15 combinations that contained either only CI or MRF, plus mixtures that contained a mixture of both. All mixtures were chosen such that they either fall within the concentration ranges in the original model or are situated on the edge of these ranges (Figure 4.1).

The model was successful at predicting the viscosity of most samples in the validation set (Figure 4.4A). The variance (RSE = 0.95) is slightly higher than the training set (RSE = 0.49), which is typical for this kind of empirical model.[154] Nevertheless, the model overpredicted the final viscosity of the samples in the validation set with higher fibre contents (Figure 4.4B). Notably, the two samples (1 and 14) with a high error contained a relatively high content of fibre isolate. Yet, not all samples with high fibre content and fibre isolate (depicted in empty symbols) resulted in a poor prediction. A comparison of the high and low residuals from mixtures with approximately the same fibre content and dry matter shows that relatively more starch improves the prediction (Figure A 4.5, Appendix). It is feasible that, at these concentrations of starch and fibre, there are other interactions between the components that are not included in the model because this area was not represented in enough detail in the training set (Figure A 4.6, Appendix). It is important to realize when making predictions based on empirical models that these predictions are solely based on the information included in the training data, without any additional mechanistic knowledge. Therefore, the predictions lose reliability when they are made for conditions outside the area defined by the training set. In the case of linear models with a single variable, the range of validity of the model can be (in most cases) defined as the range of the empirical observations. However, this task becomes much harder for empirical models with several variables, especially in

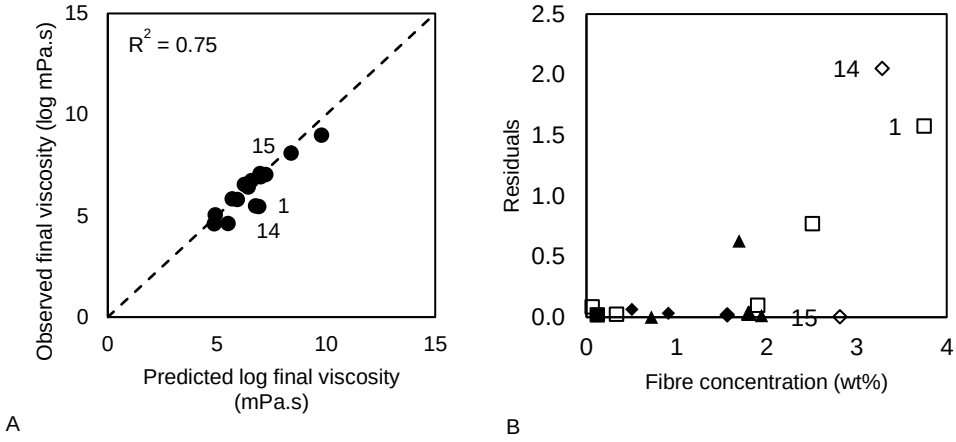

Figure 4.4 A) Parity plot of the validation of the log (base e) linear regression model with components as explanatory variables and based on the whole dataset with a) including all the validation samples and B) the fibre concentration in wt% as a function of their residuals of the prediction of the validation samples depicted for CI (■), MRF(▲), and CI-MRF mix(⧫). The empty symbols represent samples that contain fibre isolate.

those situations where interactions are relevant.

In summary, the model successfully predicts the final viscosity of the product for a broad range of concentrations, solely based on the composition (i.e. without the processing history). This highlights the potential of using observation data on functional properties to aid in product formulation. However, as is always the case with empirical models, these predictions must be used with care and, where possible, do additional experiments to prevent extrapolation beyond the measured samples. Nevertheless, 75% of the variation in the validation set was explained by the model, which is considered adequate as an illustration of these models in this study. For industrial application, we expect that the prediction should roughly indicate the level of viscosity, which helps reducing the number of application tests. This is the case for most samples in the presented model, except in the higher fibre area as indicated before, which requires more experiments.

4.4 Application

The functionality-driven selection of ingredients will now be illustrated. Our regression model shows that the viscosity obtained with functional ingredients can be described by their composition. The model for this can now be employed to reveal the range of compositions, and hence, ingredients that can be used to obtain a certain functionality in a product. The carbon footprints of these ingredients can also be considered at the same time, which allows for the selection of those formulations that have the lowest resource use. This is demonstrated for the selection of ingredients for a model system picked in this study: a salad dressing with a viscosity of 1500 mPa.s, in which the viscosity is completely determined by the techno-functional ingredients. First, the component model with a fixed intercept at 1 mPa.s was used to find all possible combinations of components that add up to 1500 mPa.s, with a variation of 2.5% allowed. In addition, a maximum concentration of 22 wt% is set and the amount of protein, starch, fibre, and rest is also constrained to what is tested in our model, to prevent extrapolation as much as possible. Evidently, when the proposed method will be used in industry, it is necessary to test all possible combinations of the ingredients at the desired concentrations.

Initially, the regression model can be solved for all combinations of components that produce a final viscosity of 1500 mPa.s. Figure 4.5 depicts the composition window in terms of ratio between the protein, starch, and fibre content that achieves this final viscosity. A viscosity of 1500 mPa.s can be achieved with a concentration ranging from 4.9 to 22 wt%. Some ratios between these three components lead to the same viscosity at multiple concentrations. This is possible since the same ratios can occur at different concentrations due to the rest component, which is not displayed in these Figures. The right ternary diagram indicates that at higher concentrations (15–22 wt%), there is a smaller composition window that results in a viscosity of 1500 mPa.s. This area contains mostly high protein blends; a higher dry matter content is required due to the relatively low contribution of protein to the overall viscosity. In the same Figure, there is a small area with a low protein, moderate fibre, and high starch content that represents the boundary of the composition that can be achieved a high starch content at high dry matter. This is most likely caused by the antagonistic effects between starch and fibre in the regression model. The use of

relatively more starch is only possible at a lower concentration. Below 15 wt%, the composition window is larger and shows that with solely fibre or starch a viscosity of 1500 mPa.s can be achieved, in contrast to protein. Both fibre and starch contribute the most to the final viscosity, and therefore, when one of these components is dominant, a low dry matter in solution is sufficient to achieve the viscosity requirement. Evidently, not all solutions in this composition window can be realized, since the solutions should be constrained by the compositions of the yellow pea ingredients to be used.

Next, all possible combinations are matched to the CI and MRF ingredients with linear optimisation using discretization on a fine grid, which led to approximately 27,000 combinations of blends (Figure 4.6). The more detailed formulations are depicted in the Appendix in Figure A 4.7. These bar charts show the concentrations of each ingredient that are used in each blend. Please note that it may seem that the bars are overlapping due to the large number of samples displayed. We can conclude that for achieving a certain final viscosity the current focus on purity in the food industry is not necessary at all: it can be achieved just as well with a wide range of blends of mildly fractionated ingredients. Of course, in practical formulation, one has to meet many more criteria than just viscosity; which will require the reduction of the formulation window by taking a cross-section with other criteria. We do expect that even then many possible combinations will remain possible even then, as there is in practice a wide choice of blends to choose from. To select the most sustainable combinations of ingredients, their global warming potential (GWP) was estimated using data from a previous life cycle assessment on the conventional and dry fractionation of yellow pea including the impact of both cultivation and processing (Table 4.1).[9] The GWP ranges from roughly 700 – 3100 kg CO_2-eq/ton product of blends and increases with the amount of CIs present (Figure 4.6). Interestingly, CIs can be used at a low GWP, which is attributed to the low GWP of starch and fibre isolate compared to protein isolate. In other words, mostly the protein isolate is responsible for the higher GWP in the blends (Figure A 4.8, Appendix). A closer look into samples on several GWP levels shows that CIs are not really necessary to for example obtain a viscous salad dressing (Figure 4.7). More specifically, combining flour and protein- and starch-rich ingredients delivers a viscosity of 1500 mPa.s at a

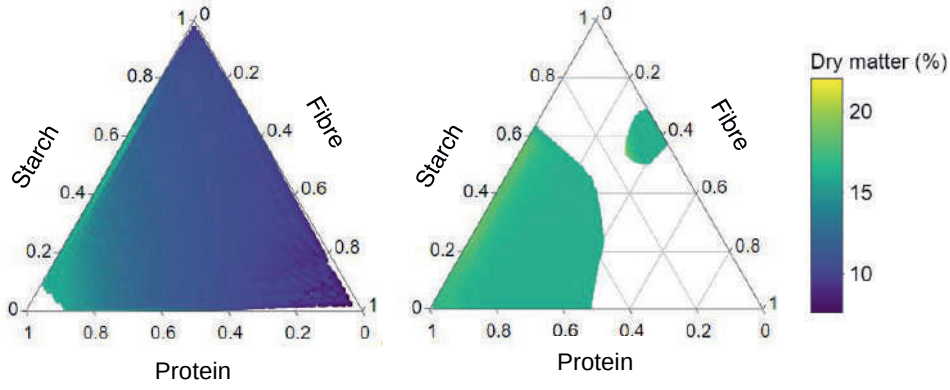

Figure 4.5 The combinations of fibre, starch, and protein that lead to a viscosity of 1,500 mPa.s, generated using the regression component model. The figure is split into two layers since some ratios (of protein, starch, fibre) lead to the same viscosity at different concentrations. The final concentration of the blend is determined by the residual fractions.

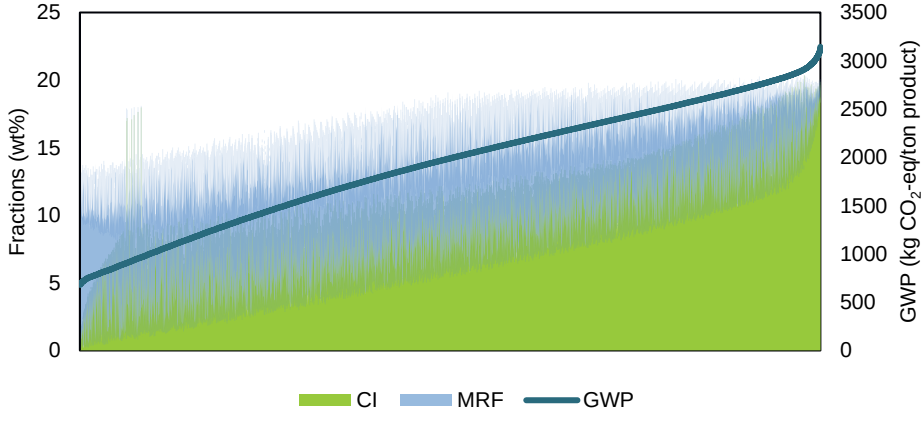

Figure 4.6 27,000 possible combinations to blend MRFs (protein- and starch-rich ingredients) with CIs (protein, starch, and fibre isolate) to obtain the viscosity of 1500 mPa.s on a discretized grid in the formulation window sorted by their global warming potential to produce a ton of product (the ingredient blend), split into CIs and MRFs.

relatively low global warming potential of 680 kg CO_2-eq/ton product of blends. In contrast, using protein isolate to thicken the dressing leads to a high GWP of up to 3143 kg CO_2-eq/ton product of blends, since protein is a poor thickening agent and therefore more is needed. On top of that, the GWP of protein is very high due to the many processing steps. Based on these results, one may suggest that protein isolates should not be used for thickening. Of course, other criteria will have to be met in practice, and there may be more other reasons to use protein, such as health. Similarly, we may prefer less refined products, which for example contain more fibre. This formulation window allows product developers to explore ingredient blends that 1) achieve a certain final viscosity required for their product, 2) meet the product requirements in terms of composition (e.g. high fibre or protein samples), and 3) have the lowest GWP. In this way, a focus on pure, standardized components such as isolates is not necessary anymore, which will enable the use of the novel, mildly fractionated ingredients. In the end, we envisage that all the important functionalities, ranging from texturizing properties (e.g. gelling, emulsification) to sensory properties (e.g. flavour, mouthfeel) can be quantified in similar ways. The product formulation then requires the identification of the cross-section of the set of blends that comply with all the criteria. One can then select one particular blend that has the lowest use of resources. This study proposes the methodology to head the towards functionality-driven selection of ingredients; in which all these functional properties can be described using data-driven methods.

4.5 Conclusions

We show with a particular case that it is possible to select food ingredients based on functionality rather than purity. This is a necessary step on the route towards the use of less refined, more resource-efficient ingredients, which will bring a strong, chain-wide reduction in resource use, while the final product functionalities will be ensured. For the illustration of functionality-driven ingredients selection, we use the final viscosity that can be achieved in formulated products with yellow pea commercial isolates and mildly refined ingredients. The commercial isolates and mildly refined ingredients from yellow pea were blended into 95 different mixtures. Starch isolate and the starch-rich ingredient from dry fractionation have a relatively strong thickening capacity compared to the protein isolate and the dry fractionated

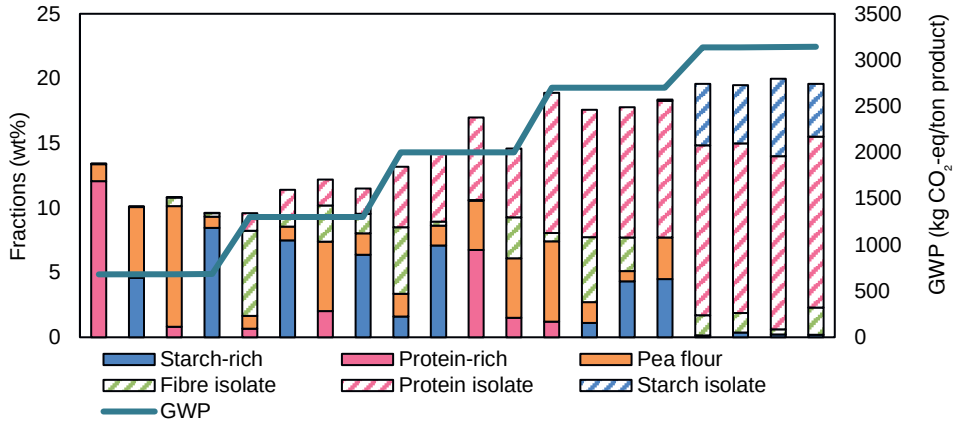

Figure 4.7 A representation of selected combinations of different GWP levels, replotted with the GWP expressed per ton product (being the final blend) on the secondary y-axis to achieve a final viscosity of 1,500 mPa.s.

protein-rich ingredient. Component-wise, starch and fibre contribute the most to the final viscosity The contribution of each component was quantified by the adapted multiple linear regression model, in which the processing history of protein and fibre did not need to be taken into account. The models were validated by an independent validation set of 15 samples. Mixtures that are high in fibre were predicted less well, which can be mainly attributed to the small number of mixtures at high fibre content in the training set. Because of this, possible interactions between components were not completely covered by the samples in the training set. The model was used to select all possible formulations for a product that should have a viscosity of 1500 mPa.s. The carbon footprint of the formulations varied from 700 to 3100 kg CO_2-eq/ton product of blends. The formulations with the higher footprint mostly consisted of commercial isolates, whereas the lower ones consisted mostly of mildly refined ingredients. We, therefore, showed that by using regression models, milder fractionated ingredients can be included to achieve the same techno-functionality as with conventional ingredients, but with a reduced global warming potential.

4.6 Appendix

Table A 4.1 Formulations of samples in wt% divided in n = 95 for the training set and n = 15 for the test set

Nr.	Coarse fraction (%)	Fine fraction (%)	Flour (%)	Fibre isolate (%)	Protein isolate (%)	Starch isolate (%)	Nr.	Coarse fraction (%)	Fine fraction (%)	Flour (%)	Fibre isolate (%)	Protein isolate (%)	Starch isolate (%)
1	8.9	0.0	0.0	0.0	0.0	0.0	56	0.0	9.8	0.0	0.0	3.4	0.0
2	9.1	0.0	0.0	0.0	0.0	0.0	57	0.0	4.0	0.0	0.0	0.0	4.9
3	11.6	0.0	0.0	0.0	0.0	0.0	58	0.0	4.0	0.0	0.0	0.0	4.9
4	11.6	0.0	0.0	0.0	0.0	0.0	59	0.0	5.0	0.0	0.0	0.0	4.4
5	4.5	0.0	0.0	0.0	0.0	0.0	60	0.0	4.8	0.0	0.0	0.0	4.3
6	5.5	0.0	0.0	0.0	0.0	0.0	61	0.0	0.0	8.9	0.0	0.0	0.0
7	4.5	0.0	0.0	0.0	0.0	0.0	62	0.0	0.0	9.1	0.0	0.0	0.0
8	5.5	0.0	0.0	0.0	0.0	0.0	63	0.0	0.0	16.7	0.0	0.0	0.0
9	5.7	0.0	0.0	0.0	0.0	0.0	64	0.0	0.0	16.7	0.0	0.0	0.0
10	3.6	0.0	0.0	5.6	0.0	0.0	65	0.0	0.0	13.0	0.0	0.0	0.0
11	3.6	0.0	0.0	5.6	0.0	0.0	66	0.0	0.0	13.0	0.0	0.0	0.0
12	6.3	2.7	0.0	0.0	0.0	0.0	67	0.0	0.0	4.6	0.0	0.0	0.0
13	6.3	2.7	0.0	0.0	0.0	0.0	68	0.0	0.0	4.6	0.0	0.0	0.0
14	6.8	2.1	0.0	0.0	0.0	0.0	69	0.0	0.0	8.0	0.0	0.0	0.0
15	6.8	2.2	0.0	0.0	0.0	0.0	70	0.0	0.0	8.0	0.0	0.0	0.0
16	9.8	3.2	0.0	0.0	0.0	0.0	71	0.0	0.0	0.0	0.0	8.9	0.0
17	9.8	3.2	0.0	0.0	0.0	0.0	72	0.0	0.0	0.0	0.0	9.0	0.0
18	0.4	10.8	0.0	0.0	0.0	0.0	73	0.0	0.0	0.0	0.0	13.8	0.0
19	0.4	10.8	0.0	0.0	0.0	0.0	74	0.0	0.0	0.0	0.0	10.8	0.0
20	0.5	12.5	0.0	0.0	0.0	0.0	75	0.0	0.0	0.0	0.0	10.9	0.0
21	0.6	12.8	0.0	0.0	0.0	0.0	76	0.0	0.0	0.0	0.0	13.8	0.0
22	0.6	11.4	0.0	0.0	0.0	0.0	77	0.0	0.0	0.0	0.0	4.1	4.9
23	0.6	11.5	0.0	0.0	0.0	0.0	78	0.0	0.0	0.0	0.0	4.1	4.9
24	6.3	0.0	0.0	0.0	2.8	0.0	79	0.0	0.0	0.0	0.0	10.9	2.5
25	6.3	0.0	0.0	0.0	2.8	0.0	80	0.0	0.0	0.0	0.0	10.9	2.5
26	0.0	0.0	0.0	9.3	0.0	0.0	81	0.0	0.0	0.0	0.0	2.9	4.4
27	0.0	0.0	0.0	9.3	0.0	0.0	82	0.0	0.0	0.0	0.0	3.0	4.5
28	0.0	9.8	0.0	0.9	0.0	0.0	83	0.0	0.0	0.0	7.2	0.0	5.9
29	0.0	9.9	0.0	0.9	0.0	0.0	84	0.0	0.0	0.0	6.5	0.0	6.5
30	0.0	11.8	0.0	0.9	0.0	0.0	85	0.0	0.0	0.0	4.6	0.0	8.5
31	0.0	11.8	0.0	0.9	0.0	0.0	86	0.0	0.0	0.0	2.6	0.0	10.4
32	0.0	0.0	0.0	10.6	10.9	0.0	87	0.0	0.0	0.0	0.0	0.0	8.8
33	0.0	0.0	0.0	10.6	10.8	0.0	88	0.0	0.0	0.0	0.0	0.0	9.1
34	0.0	0.0	0.0	6.8	0.0	2.3	89	0.0	0.0	0.0	0.0	0.0	4.4
35	0.0	0.0	0.0	6.9	0.0	2.3	90	0.0	0.0	0.0	0.0	0.0	4.4
36	0.0	8.8	0.0	0.0	0.0	0.0	91	0.0	0.0	0.0	0.0	0.0	4.4
37	0.0	9.0	0.0	0.0	0.0	0.0	92	0.0	0.0	0.0	0.0	0.0	6.7
38	0.0	13.7	0.0	0.0	0.0	0.0	93	0.0	0.0	0.0	0.0	0.0	6.7
39	0.0	22.8	0.0	0.0	0.0	0.0	94	0.0	0.0	0.0	0.0	0.0	2.5
40	0.0	22.8	0.0	0.0	0.0	0.0	95	0.0	0.0	0.0	0.0	0.0	2.5
41	0.0	18.2	0.0	0.0	0.0	0.0	1.test	0.0	0.0	0.0	7.0	0.0	0.0
42	0.0	18.3	0.0	0.0	0.0	0.0	2.test	0.0	0.0	0.0	0.0	15.0	0.0
43	0.0	13.7	0.0	0.0	0.0	0.0	3.test	0.0	0.0	0.0	0.0	0.0	8.0
44	0.0	13.7	0.0	0.0	0.0	0.0	4.test	0.0	0.0	0.0	3.5	0.0	3.5
45	0.0	10.5	0.6	0.0	0.0	0.0	5.test	0.0	0.0	0.0	4.5	4.5	0.0
46	0.0	10.6	0.5	0.0	0.0	0.0	6.test	0.0	0.0	0.0	0.0	4.0	4.0
47	0.0	12.7	0.7	0.0	0.0	0.0	7.test	8.0	0.0	0.0	0.0	0.0	0.0
48	0.0	12.6	0.8	0.0	0.0	0.0	8.test	0.0	10.0	0.0	0.0	0.0	0.0
49	0.0	11.4	0.8	0.0	0.0	0.0	9.test	0.0	0.0	16.0	0.0	0.0	0.0
50	0.0	11.4	0.7	0.0	0.0	0.0	10.test	5.5	5.5	0.0	0.0	0.0	0.0
51	0.0	2.0	0.0	0.0	7.2	0.0	11.test	0.0	6.0	6.0	0.0	0.0	0.0
52	0.0	2.0	0.0	0.0	7.2	0.0	12.test	0.0	4.5	0.0	0.0	0.0	4.5
53	0.0	3.0	0.0	0.0	10.1	0.0	13.test	4.5	0.0	0.0	0.0	4.5	0.0
54	0.0	3.0	0.0	0.0	10.2	0.0	14.test	0.0	4.5	0.0	4.5	0.0	0.0
55	0.0	9.7	0.0	0.0	3.3	0.0	15.test	4.5	0.0	0.0	4.5	0.0	0.0

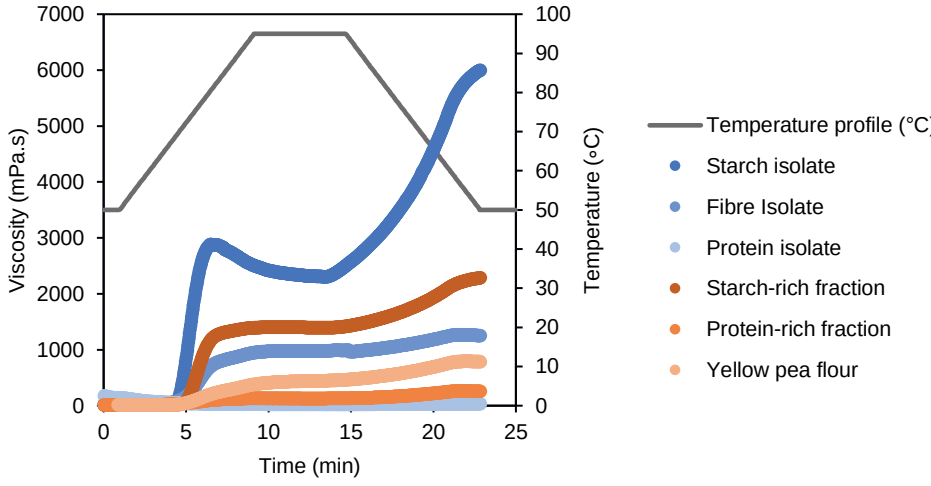

Figure A 4.1 Temperature - time profile for the determination of the viscosity including the viscosity profiles for all yellow pea ingredients used in this study: starch, fibre, and protein isolate, and the milder fractionated starch-and protein rich fraction as well as the yellow pea flour. The ingredients were measured with a dry matter content of 10wt%.

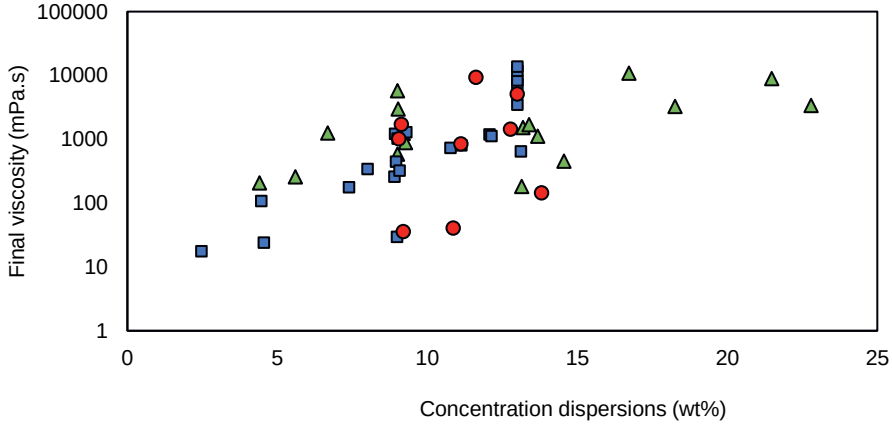

Figure A 4.2 The ingredient contribution to the final viscosity plotted over dry matter. Here the samples are grouped into only isolates, only mildly refined and only mixtures. The symbols indicate (●) mixtures of isolates and MRFs, (■) MRFs single and mixed, (▲).

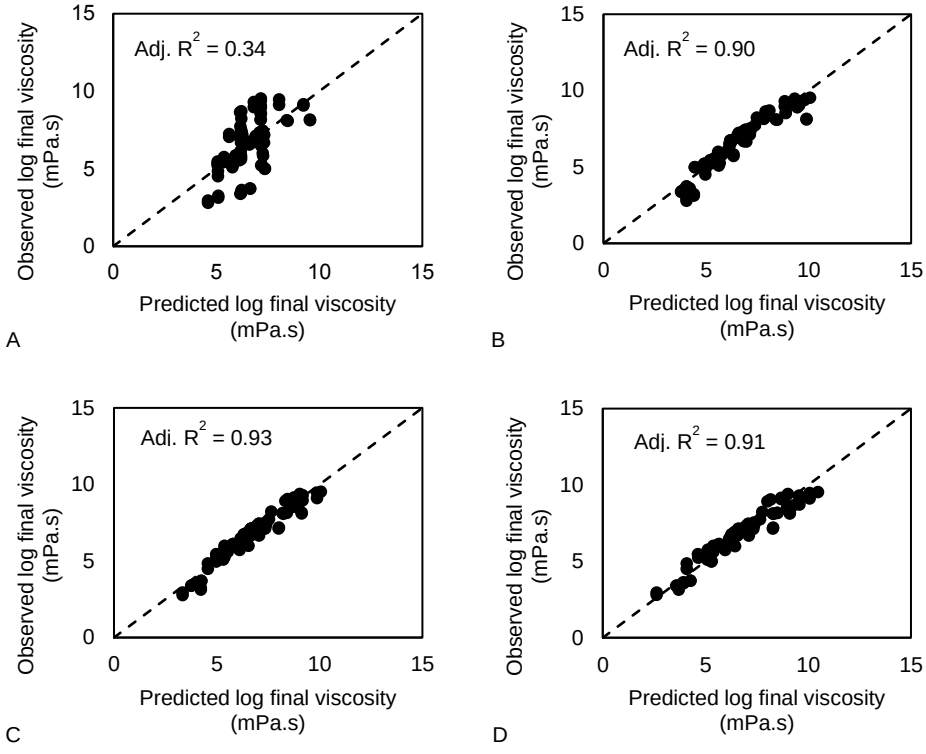

Figure A 4.3 Observed versus predicted log (with base e) final viscosity in mPa.s by a) dry matter b) ingredients c) component and d) components with a forced intercept through 1 mPa.s. The dashed diagonal line indicates the samples that would have a perfect fit.

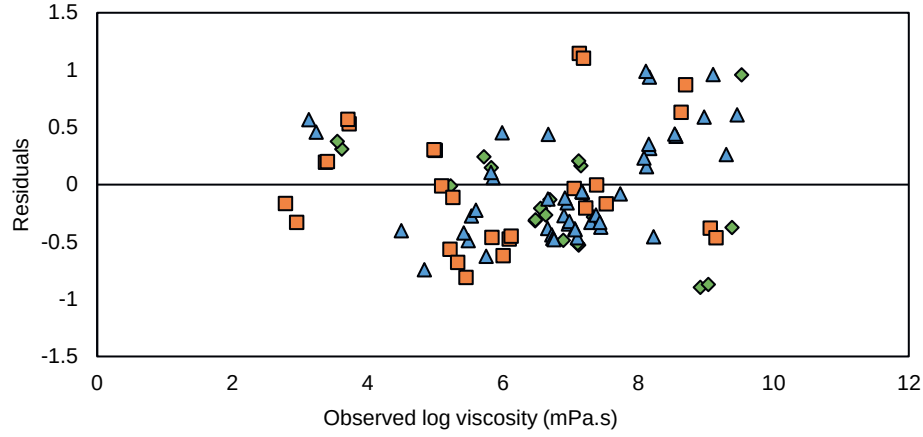

Figure A 4.4 Residuals of the log (base e) component model with intercept fixed at 1 mPa.s, split for CIs (■), MRFs (▲), and mixtures (♦) as a function of the observed log (base e) final viscosity.

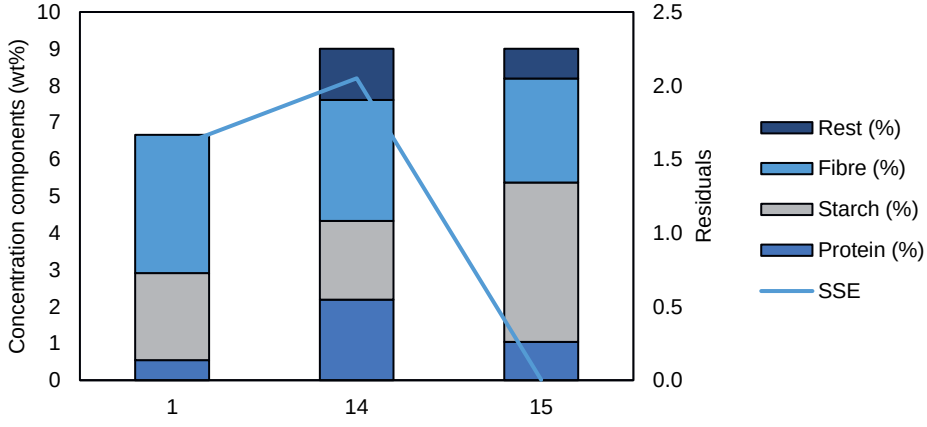

Figure A 4.5 Comparison of selected mixtures on composition and their residual value. The sample numbers on the x-axis are the same as in Figure 4.

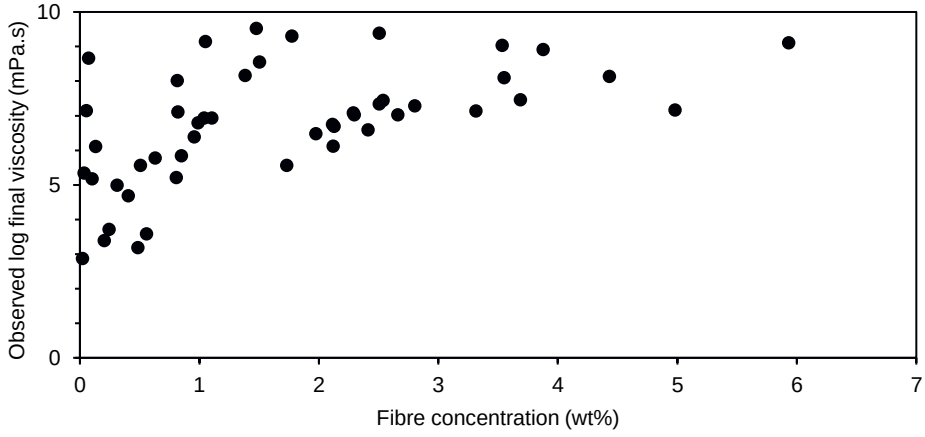

Figure A 4.6 Fibre content of all samples evaluated in the training set in dry matter % as a function of the log (with base e) final viscosity in mPa.s.

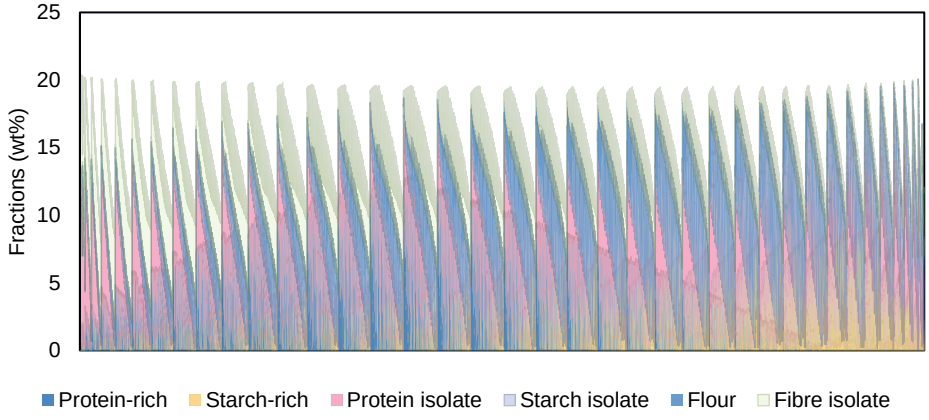

Figure A 4.7 All possible combinations to blend MRFs (protein- and starch-rich ingredients and flour) with CIs (protein, starch, and fibre isolate) to obtain the viscosity of 1500 mPa.s on a discretized grid in the formulation window.

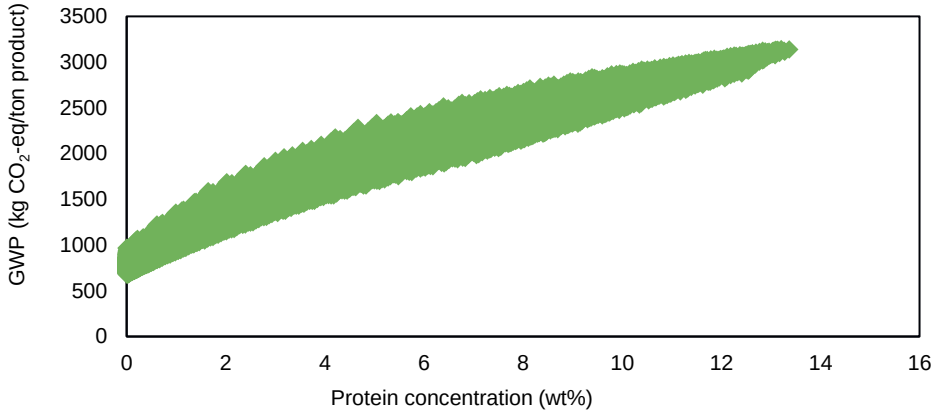

Figure A 4.8 Protein concentrations (wt%) in selected blends in plotted with their kg CO_2-eq/ ton product of ingredient blend).

Chapter 5

Framework

Machine learning to quantify techno-functional properties – A case study for gel stiffness with pea ingredients

---

A. Lie-Piang, A. Garre, T. Nissink, N. van Beek, A. van der Padt,

R.M. Boom

Innovative Food Science and Technology (2023)

Abstract

Mildly refined ingredients are included more easily in food products when selected based on techno-functional properties instead of composition. We assess different machine learning methods that quantitatively link relevant techno-functional properties to the composition and processing history of the ingredient in a case study using the gel stiffness (Young's modulus) by conventionally and mildly refined ingredients of yellow pea. Linear (multiple, log transformed and polynomial) and non-linear models (spline regression, decision trees, and neural networks) were explored. The final model selection was based on 1) the statistical model metrics (RMSE, R^2, and MAE) of the training and independent test set and 2) expert knowledge to evaluate the plausibility of the model predictions. In this case, neural networks can describe the gel stiffness of yellow pea ingredients most accurately. The approach that we follow can be applied to other techno-functional properties to improve the chain sustainability while ensuring the full functionality of the products.

5.1 Introduction

Conventional ingredients to formulate and texturize food products are highly purified and have known and standardized techno-functional properties.[2] Their isolation is intensive in the use of raw materials, water and energy and results in non-food grade, low-quality side streams.[3] Mildly refined variants of ingredients are often overlooked or not considered due to their complex multi-functional behaviour. Nevertheless, less refined ingredients inherently require fewer resources to be produced; in addition, they can be suitable for a 'clean' label and generally are more sustainable.[8,9] It is proposed that to use these ingredients, the focus on purity should be shifted to their textural contribution to the final food product.[2]

In our previous study, we showed that food products can be reformulated based on targeted textural properties, in that case the final viscosity, while controlling the nutritional qualities and minimizing resource use.[129] It was found that in such functionality-driven food product formulation, milder variants of textural ingredients can be included easily, regardless of the complex compositions. The relationship between the composition of the ingredients and viscosity could be quantified using linear regression. Such data-driven models can relate an output variable to many input variables and their interactions. In product design there are many more attributes to consider, such as taste and other techno-functional properties. Therefore, a logical next step is to quantify the gelling capacity of yellow pea ingredients.

For the gelation of ingredients similar mechanisms as viscosity are at stake: starch gelatinisation[66], protein denaturation and aggregation[158], and water uptake of dietary fibres[81]. Kornet et al.[99] found that a more mildly fractionated yellow pea protein fraction produced without isoelectric precipitation leads to stronger gels than the precipitated protein isolate. In contrast, Pelgrom et al.[66] found that protein isolate gels are stronger than gels made with dry fractionated protein-rich ingredients or with an ultra-filtrated version of the latter. These studies highlight the complexity to predict gelling properties of ingredients with a different processing history, which is a property that probably depends non-linearly on the composition and possibly on other parameters.

Non-linear relationships can be described by for example non-linear explicit equations or by machine learning algorithms. The latter has been used already in this field to predict deviations in food waste production[26], to predict the extendibility of food materials during 3D printing[159], and to model the effect of process conditions on the textural properties of non-fat yoghurt[130]. These algorithms are better suited for establishing adequate predictions rather than for gaining insight into the underlying chemical/physical phenomena. In addition, these methods can incorporate many independent variables. Therefore, they could be useful in modelling complex non-linear techno-functional properties such as taste and other textural properties, to identify and quantify a functionality-based formulation window for food products.

Although these models can often describe complex datasets, they can also introduce empirical artefacts that could even violate basic physical laws (e.g., unreasonable maxima in stiffness). Besides, there is a large number of available Machine Learning algorithms. Therefore, we propose in this study a novel standardized approach for model selection that combines both statistical aspects and expert knowledge regarding the feasibility of the model predictions. The aim of this study is to develop a procedure to select appropriate machine learning methods to quantify complex functional properties, illustrated using the gel stiffness of yellow pea ingredients. The explored models range from relatively simple linear regression to more complex neural networks. Since appropriate data is key to creating any model, initially, the experimental data of the gel stiffness from conventionally (fully) and mildly refined ingredients are obtained and evaluated. After that, the approach to select an appropriate machine learning model is applied to the data and suitable models are selected. The acceptability of a model is derived from the model metrics and the constraint that the behaviour is physically feasible.

5.2 Materials & methods

5.2.1 Materials

Pre-dried yellow peas (*Pisum sativum L.*) were purchased from Alimex (The Netherlands). Commercial pea protein isolate (Nutralys F85G) and pea fibre isolate (PEA FIBRE I 50 M) were purchased from Roquette Frères S.A (St. Louis, USA). Pea

starch isolate was purchased from Emsland Stärke GmbH (Germany). The water used in all experiments was deionized.

5.2.1.1 Production of mildly refined ingredients

The hulls were initially removed from the yellow peas with a Satake TM05 pearling machine (Satake Corporation, Japan). The dehulled peas were then pre-milled into grits using a pin mill (LV 15M, Condux-Werk, Germany) at room temperature. Milling and air classification took place in a room with a controlled relative humidity of 30%. The grits were subsequently milled into a dehulled fine flour (PF) using a ZPS50 impact mill (Hookah-Alpine, Germany), with an impact mill speed of 8000 pm, the airflow at 52 m^3/h, the classifier wheel speed at 4000 pm, and a feed rate of 2 rpm (method adopted[49]). A thermometer inside the mill indicated a temperature between 16 and 34°C. Yellow pea ingredients enriched in protein (PR) and starch (SR) were obtained by air classification. An ATP50 classifier (Hosokawa-Alpine, Germany) was used, with a fixed airflow at 52 m^3/h, the classifier wheel speed at 5000 rpm, and the feed rate at 20 rpm.

5.2.2 Compositional analysis

The protein content of the ingredients was determined with the Kjedahl method; a nitrogen conversion factor of 5.52 for pea protein was used.[137] The total starch content was determined with a Total Starch Amyloglucosidase-α-Amylase Assay Kit, AOAC Method 996.11 (Megazyme International Ireland Ltd, Bray, Ireland). The total dietary fibre content was determined using the AOAC method 985.29. The moisture content was determined using a vacuum oven at 70°C. The rest of the components were determined as the difference between the dry matter and the protein, starch, and fibre content and contain among others oil, ash, salts, and sugars (Table 5.1). All protein in the isolates is considered denatured[10] and all protein in the mildly refined fractions is considered native.

The nitrogen solubility index (NSI%) was determined using an adapted protocol.[160] A 1 wt% dispersion of each ingredient was thoroughly mixed and rotated for 30 minutes. The dispersions were then centrifuged (30 min, 10,000 g, at 20°C). The supernatant was removed, after which the weight of the pellet was determined in the

tube, to prevent mass loss. Next, the pellet was dried in an oven at 105 °C for 24 hours and the dry matter content was determined. The protein content of the pellet was also determined using the aforementioned Dumas method, which was corrected for any extra water uptake between the preparation of samples and measurement in the Dumas. The NSI% was calculated by the percentage of soluble protein of the total protein present in the ingredients.

5.2.3 Gel formulation & preparation

The conventional ingredients (protein, starch, and fibre isolates) and the dry fractionated ingredients (pea flour and protein- and starch-rich) were mixed into different formulations. The formulations space was designed such that the whole formulation window by these ingredients is covered, in all possible combinations of:

- pure isolates and mixtures of those,
- mildly refined ingredients and mixtures of those
- mixtures of pure isolates and mildly refined ingredients.

Table 5.1 Composition of ingredients based on g/100 g dry matter and nitrogen solubility index, with mean and standard deviation. Two batches of flour were used, one for the gelation experiments and one for the viscosity experiments.

	Protein	+/-	Starch	+/-	Fibre	+/-	Rest	+/-	Solids	+/-	NSI%	+/-
Protein isolate	74.1	1.8	0.4	0.03	2.2	1.2	23.3	1.6	93.7	0.5	16.4	5.8
Starch isolate	0.5	0.0	91.4	2.9	0.8	0.8	7.3	0.0	88.4	0.5	0.0	0.0
Fibre isolate	7.8	0.5	33.8	0.5	53.6	4.6	4.8	1.1	95.2	0.5	34.7	9.0
Protein-rich fraction	44.4	1.1	6.9	0.2	18.9	2.9	29.8	2.1	93.5	0.5	48.0	0.8
Starch-rich Fraction	10.5	0.5	66.6	1.9	4.6	1.5	18.3	1.6	90.9	0.5	55.3	7.5
Flour viscosity	22.3	0.8	47.9	7.1	10.6	0.1	19.2	0.4	90.6	0.5	61.3	0.9
Flour gel	25.2	0.7	48.2	0.5	9.3	2.1	17.4	0.0	90.6	0.5	61.3	0.9

The selection of samples ensured that a formulation window from high to low starch, protein, and fibre concentrations was covered (Figure 5.1). Since this formulation window is constrained by the composition of the individual ingredients, no formulations are present with more than 53.5 wt% fibre, which is the highest fibre concentration practically possible with these ingredients (Table 5.1). The formulations were divided into a train and test set, which is explained in section 5.2.4. The train and test formulations were created at concentrations 13, 18, and 24 wt%. In addition, interpolation of the models was tested with additional formulations at 15 and 21 wt%. This resulted in a total number of (n = 91) train and (n = 72) test formulations.

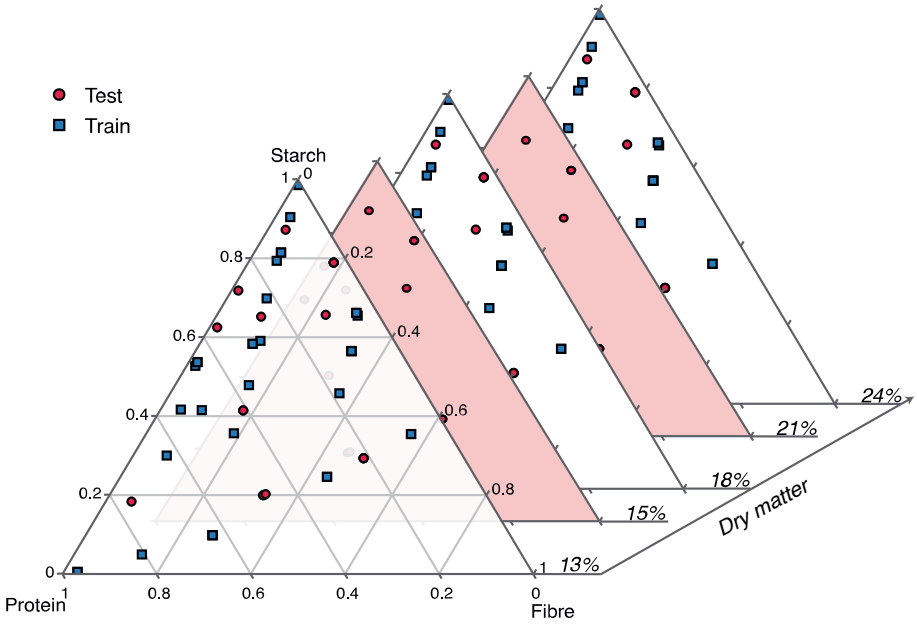

Figure 5.1 Ternary diagram of training and test set gels, created by mixing the commercial isolates and mildly refined ingredients. The diagram depicts the ratio between the starch, protein, and fibre content, please note that there is also a rest fraction that is not shown. All ternary diagrams depict the same samples, yet in different concentrations. The white ternary plots indicate concentrations trained and tested for (13, 18, 25%) and the red ternary diagrams indicate concentrations at which only test set gels are measured (15 and 21%).

The gels were prepared using an adapted protocol.[161] The dry ingredients were first premixed according to the formulations mentioned above and subsequently dispersed in deionized water under mild agitation using a magnetic stirrer for 5 to 10 min. The duration of stirring depended on the consistency of the dispersion. Where necessary, manual stirring was added to enhance the dispersion. The mixtures were left to hydrate while still stirring gently for one hour at room temperature covered with parafilm. They were subsequently transferred to in-house produced Teflon tubes with a diameter of 20 mm and length of 100 mm. The tubes with the mixtures were heated in a water bath at 90 °C for 30 min in a rotator at a speed of 90 Hz. They were then removed from the water bath and left for one hour to cool down whilst remaining at the rotation speed to prevent any possible sedimentation. Next, the tubes were stored upright in the fridge at 4 °C overnight. The tubes were removed from the fridge one hour before further measurements. The gels were carefully released from the tube and cut into 3 x 20 mm high cylinders with an in-house produced knife. All three cylinders were measured when possible.

The fracture stress and strain were measured by uni-axial compression with a Texture Analyser (Instron-5564 Series Table Model Systems Twin-column design, Canton, USA) using a 100 N load cell and a 5.08 cm cylindrical pressure head. A 90 % deformation compression test at 1 mm/s was performed on three slices from each gel. All measurements were at room temperature. The compressive strain and stress were recalculated to the true Henky's stress and strain according to another study.[161] The true Henky's stress (ε_h) is defined in Equation 5-1 below where h_0 is the initial height of the gel and $h(t)$ is the height after deformation time t. The true Henky's strain is negative for compression but is expressed as an absolute value.

$$\varepsilon_h = \ln\left(\frac{h(t)}{h_0}\right) \tag{5-1}$$

To calculate the true Henky's stress, the initial surface contact area A_0 $[m^2]$ is corrected for the change during compression while assuming the volume stays the same using Equation 5-2.

$$A(t) = \frac{h_0}{h(t)} A_0 \tag{5-2}$$

Using the corrected surface area, the true Henky's stress σ $[Pa]$ was calculated with Equation 5-3 in which $F(t)$ is the force per unit of area $A(t)$.

$$\sigma(t) = \frac{F(t)}{A(t)} \tag{5-3}$$

Using the true Henky's strain and stress, the Young's modulus (E_u) [Pa] was calculated using Equation 5-4 and always from the linear part within the region 0.02 - 0.15 of the fracture strain.

$$E_u = \left(\frac{d\sigma}{d\epsilon_h}\right) \tag{5-4}$$

5.2.4 Model selection

All models were fitted with RStudio version 4.0.1 using the adopted step-wise linear regression as done by Lie-Piang et al.[129] or with the package caret version 6.0.92[162] The total dataset was split into a train (56 %) and an independent test set (44%). We used a 5-time repeated 10-fold cross-validation using the training set to minimize overfitting and determine the values for the parameters of the models, which are also called hyperparameters. These parameters define the architecture of the models (e.g. number of nodes in a neural network)[163] and are chosen based on the lowest cross-validation error. An example of tuning plots for hyperparameters is provided in the Appendix (Figure A 5.2). The test set was used independently to assess the performance of the model. The complete code of the analysis is available in Open Code from the GitWUR page of the main author: https://git.wur.nl/anouk.lie-piang/toward-ml-for-food-formulation.

The approach to select the appropriate machine learning method is summarized in Figure 5.2 and consists of three areas: 1) Creating different types of models with different sets of variables 2) assessing the feasibility of the model predictions and 3) evaluating model metrics. Each of these steps will be discussed in detail in the following sections.

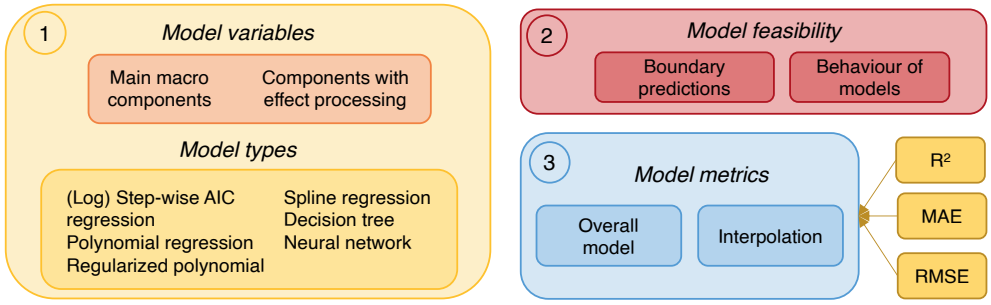

Figure 5.2 Proposed approach to select suitable models for quantifying functional properties.

5.2.4.1 Model variables and types

In every model, the dependent variable (Y) is the gel stiffness expressed in the Young's modulus. Several sets of input variables X_i were evaluated. The first set of variables contained the main macro components (in wt%), with the index i representing protein, starch, fibre, and the rest of the components that were considered as one fraction. In the other two sets of variables, the effect of the processing history was included based on the nativity and solubility of the protein in the ingredients. Hence, i includes now not the overall protein composition, but a separate entry for the soluble and insoluble protein or the native and denatured protein, both in wt%, in combination with the other components starch, fibre and rest.

Multiple linear regression with a stepwise selection based on the AIC criteria was adopted from Lie-Piang et al.[129] The interaction terms between variables X_i are specifically included in this model. Further, a log-transformed dependent variable Y was also used in linear regression. Polynomial regression was also applied, in which the independent variables are raised to the n^{th} degree, with n = 1, 2, 3, 4.[164] The degree n is the hyperparameter that is determined by the lowest test error. Regularisation was further applied to the polynomial regression to reduce the number of parameters by introducing a penalty function in the calculation of the cost function. Three commonly used regularisation methods were considered: ridge regression, lasso regression, and elastic net. Ridge regression adds a penalty (λ) in the calculation of the Sum of Squares defined by $\lambda \cdot \sum_{i=1}^{n} \beta_i^2$ and thereby forcing

coefficients (β_i) to become smaller. Lasso regression uses the absolute values of the coefficient: $\lambda \cdot \sum_{i=1}^{n} |\beta_i^i|$, and can thereby force irrelevant coefficients to zero. Elastic net regression is a combination of Lasso and Ridge regression, where the penalty function has two λ parameters.[165] The hyperparameter α, which is 1 for Lasso regression, 0 for Ridge regression, and in between for elastic net in $\lambda \cdot \left[(1-\alpha)\sum_{i=1}^{n}\beta_i^2 + \alpha\sum_{i=1}^{n}|\beta_i^i|\right]$.

Spline regression was also evaluated, which is a non-parametric regression technique that recursively partitions the data into splines (regions), in which each is fitted by local least squares.[166] The model's hyperparameters can be tuned for smoothness and flexibility by controlling the number of regions, or knots. Next, a random forest was applied, which is based on bootstrapped samples from the dataset combined with a random selection of the input variables X_i used to describe each sample, reducing parameter correlation.[167] The number of variables to be finally used is the hyperparameter that is tuned.

Finally, a neural network was fitted to the data, which has several layers with "hidden nodes" between the model inputs and outputs.[168] The model was fitted using the backpropagation algorithm. The combination of inputs for each neuron was defined by the Sigmoid function $\sigma(v) = 1/(1+e^{-v})$, where v is the output of the linear combination. For this study, softplus, ReLu and sigmoid activation functions were compared without observing major differences, therefore we used the latter. The input X_i and output variable Y were scaled and centred (training and independent test set gels separately) since this reduced the computing time of a neural network. The optimal number of layers (maximum two to prevent overfitting) and nodes (maximum three) in those layers are the hyperparameters that were tuned.

5.2.4.1.1 Significant differences between sets of variables

The models were subjected to k-fold cross-validation and the neural networks were run five times to prove significant differences between the sets of variables. To reveal differences between the sets of variables, a Two-Way ANOVA and Tukey's post-hoc test were performed, with $p \leq 0.05$. A repetition of five times was found sufficient,

as more (10 or 20 times) did not change the outcome significantly (95% CI with $p \leq 0.05$).

5.2.4.2 Model feasibility

Each model was evaluated for its feasibility in the prediction of values close to the boundary values. In the case of gel stiffness, these were the gels with lower Young's modulus values and gels that did not form a gel at all. Moreover, we require that the models should have physically plausible behaviour at all potential formulations. In other words, no unrealistic peaks or dips should be generated that indicate the presence of artefacts in the models. Therefore, the physical feasibility of the behaviour of the models was plotted as a function of the independent variables. The models were applied to a theoretical dataset in which each component varied from 1-15 wt%, whereas the others were fixed at 2 wt%. Notably, scaled Y_{scale} values for neural networks were back-transformed with the mean and standard deviation from the training set with $Y = Y_{scale}/sd + mean$, with sd being the standard deviation.

5.2.4.3 Model metrics evaluation

The model metrics were evaluated for the complete dataset and the test sets at 15 and 21 wt% separately. Metrics evaluated were the coefficient of determination R^2 based on the squared correlation, the Root Mean Squared Error ($RMSE$) based on: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and Mean Absolute Error (MAE) with: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ with observed y_i and calculated values $\hat{y}_i$ and $\bar{y}$ the mean of y_i.[162] Please note that the R^2 solely provides information on the correlation between the predicted and observed values, therefore, the error in terms of RMSE and MAE should always be considered at the same time. In addition, the model metrics for the log scale linear regression model and the neural network are based on the back-transformed data to ensure a fair comparison between the different models.[169]

5.3 Results & discussion

5.3.1 Experimental data

The basis of any machine learning model is a suitable set of experimental data points that relate the independent variables to the dependent variable. Therefore, a protein, starch, and fibre isolate were mixed in different combinations with a mildly refined protein- and starch-rich fraction. The behaviour of the single ingredients can be found in the Appendix (Figure A 5.1).

The Young's modulus of all analysed gels in the range of 13-25 wt% is plotted in Figure 5.3. A rough trend is visible between the starch concentration and the Young's modulus. The gel stiffness generally decreases with more protein; for fibre the trend is less clear, as it depends mainly on the rest of the components present. Similar observations were found by Pelgrom et al.[66] who created gels from conventionally and mildly refined ingredients and found an exponential increase in gel strength with the starch concentration and a decrease in the protein and fibre concentrations. However, this relation between protein and fibre is less evident compared to starch, which indicates that interactions with other components will be relevant as well. To support the functionality-driven product formulation, mathematical models that account for such a (potentially complex) interactions are required.

5.3.2 Selection of models to predict gel stiffness

5.3.2.1 Model variables and types for gel stiffness

All model types described in the material and methods section are evaluated, namely, linear regression, with and without a log-transformed *Y*, polynomial regression and a regularized version of polynomial regression, spline regression, random forest, and neural network. For these models, three sets of variables were considered. The first set includes the main macro components protein, starch, fibre, and the rest of the components. In the other two sets of variables, the processing history is also considered (i.e., differences between mild and extensive refining). All hyperparameters that were tuned for both functional properties and sets of variables can be found in the Appendix (Table A 5.1).

The effect of the processing history, in general, has been found especially important for the functionality of the protein. The processing history is reflected in the degree of denaturation, induced during alkaline extraction with isoelectric precipitation [10] and/or the spray drying step[170]. However, the difference in gelation behaviour is not only caused by the nativity of proteins as such. During the production of commercial isolates, the precipitation of the proteins at the isoelectric point yields mostly globulins, while the proteins that do not precipitate (mostly albumins) are discarded and lost.[10,64] These two types of protein result in different gel strengths.[94,171] This means that soluble (albumins) and insoluble (globulins) protein present in either the commercial isolates or mildly refined ingredients may behave differently. Therefore, the second and third sets of variables use split protein entries according to their solubility (NSI%) and nativity.

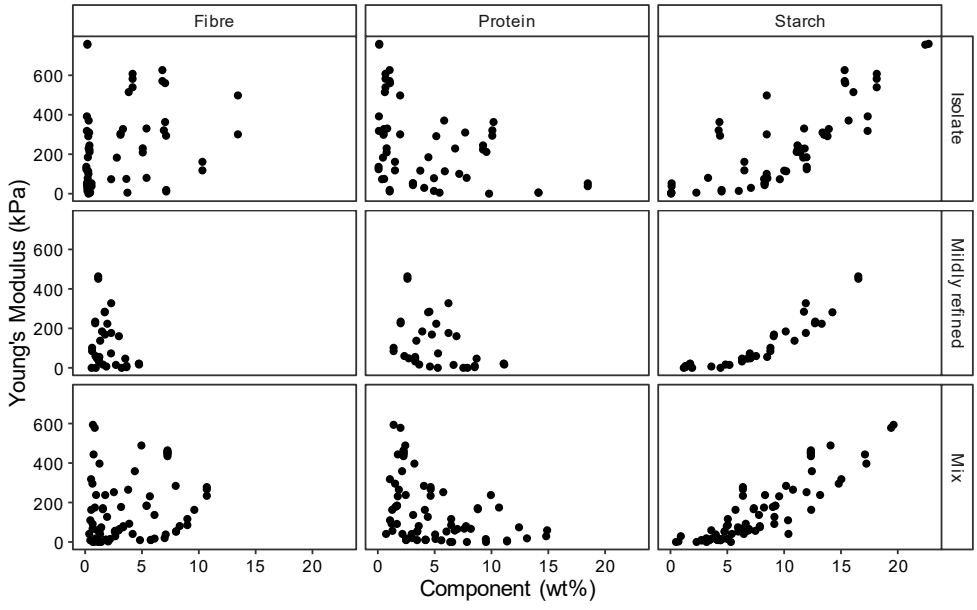

Figure 5.3 Observed Young's modulus of isolates, mildly refined ingredients and mixtures of both the isolates and the mildly refined ingredients (mix), plotted as a function of component concentration (wt%).

5.3.2.2 Model feasibility for gel stiffness

5.3.2.2.1 Feasibility of model predictions for gels with low stiffness

For the functionality-driven formulation of foods it is especially important that models can adequately describe the values close to both the higher and lower boundaries of the gel stiffness. Since it is mostly the gels with a low stiffness that were poorly predicted, we only discuss these here. In addition, some dispersions did not gel at all. This is evaluated subsequently. The conclusions for all three sets of variables are similar – therefore, we only show below the results of the set of variables with the macro components (other results can be found in the Appendix).

The models that predict negative values are plotted in Figure 5.4A. The linear and polynomial regression as well as the regularisation, predicted many negative values of Young's modulus in the test and training set. Naturally, the log-linear model does not predict negative values due to the log transformation. The spline regression and neural network can distinguish between gels with low and high Young's moduli, except for a few samples, which is negligible relative to the size of the dataset. The random forest uses another type of splitting, namely recursive partitioning and therefore can distinguish between large ranges of values as well.

The suitability and practical applicability of a model are higher if it can also identify formulations that do not gel at all (i.e., Young's modulus of 0 kPa), by giving it either a very low value or even 0 kPa. Figure 5.4B shows that none of the models assigns 0 kPa to all gels with low gel strength. Only the log-linear model, random forest, and neural network come relatively close to 0 kPa in the training set, but do not produce the same behaviour in the test set. Interestingly, the spline model predicts the dispersions that did not gel (n = 0) only at two Young's modulus values of 15 and 22 kPa. This indicates that the spline regression is able to group dispersions that do not gel, albeit returning a slightly higher Young's modulus than 0 kPa. The reason for the poor prediction at 0 kPa value could be because this area is undertrained. Nevertheless, the gel stiffness ranges between 0 and 759 kPa, and values between -50 and +50 are probably insignificant in any formulation aimed at gelation, as 50 kPa is only 6.5% of the maximum gel strength.

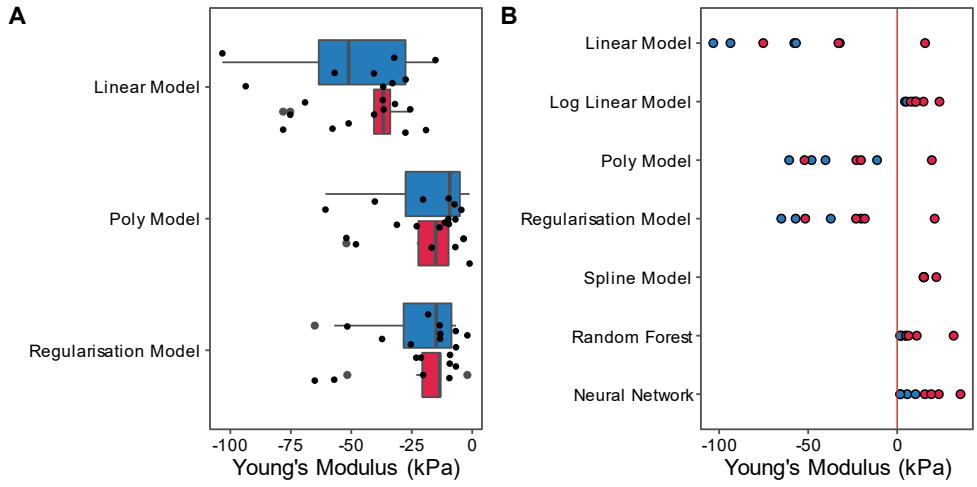

Figure 5.4 Feasibility of gels at the lower boundary with model based on main macro components. A) Distribution of the data points for which model fits (training set in blue) or predicts (test set in red) a negative value B) model fits (training in blue) or prediction (test in red) for gels that did not gel (0 kPa). Please note that in B some samples are overlaying due to a similar prediction, but for all models the same number of samples is plotted.

5.3.2.2.2 Physical feasibility of the model behaviour

Since the models will be used to formulate food ingredients for a range of compositions, it is important to evaluate the model predictions within the formulation window. The main purpose of this analysis is to identify unrealistic inflection points and discontinuities that may be an artefact of the fitting method and could compromise future applications of the models. The aim of this analysis is not to assess the accuracy of the predictions, which will be done in a later section of this paper using parity plots and model accuracy metrics.

Figure 5.5 shows the behaviour of the models trained with the main macro components as independent variables. The protein, starch, and fibre concentrations individually vary between 0-15 wt%, while the other components are fixed at 2 wt%. Starch is clearly the largest contributor to the gel stiffness, which was also observed in the experimental data (Figure 5.3). Moreover, the differences between the types of models are clear from the Figure:

1) The linear relation is, evidently, a straight line, which is the cause of the predicted negative gel stiffness values.

2) The linear regression on the log scale shows the expected exponential curve that cannot go negative.

3) The polynomial model is raised to the third degree (Appendix Figure A 5.2 and Table A 5.1). However, slightly negative values are also predicted, which is caused by the parameters raised to the first degree. Parameter reduction with regularisation did not solve this matter, as there was only a slight difference with the original polynomial regression.

4) In spline regression, the splines are very clear; there is an increase in gel strength at 5 wt% concentration.

5) The recursive partitioning in the random forest is also clearly visualized in the Figure, in which there is a discontinuous increase in the values of Young's Modulus. A discontinuous behaviour like this could pose challenges for the later optimisation of formulations and is therefore not preferred.

6) The neural network does show a continuous response, which increases with component concentration. As the trend of the neural network is similar to the trend of the linear regression on the log scale, the neural network can recognize the exponential behaviour in gel strength.

These six observations also hold for the models trained with the sets of variables with the protein split (Figure 5.6). Due to the similarity of the results, only the Figures for the protein split according to solubility are shown. The information for the split of protein according to nativity can be found in the Appendix. Not all models return a different behaviour for soluble and insoluble protein. Interestingly, only the linear regression considers soluble protein as a somewhat more important component. Based on the experimental data, protein does at some point have an important contribution, which might be linked to their solubility. The reason why this behaviour is only present in this model is not very clear. This highlights that machine learning types of algorithms are rather a tool for obtaining proper predictions rather

than understanding behaviours. Nevertheless, the models give an indication for plausible behaviours, which can subsequently be further investigated.

In summary, the linear, polynomial, and regularisation regression poorly predict the gel stiffness near the lower boundary. The feasibility study also shows that the random forest does give a continuous curve and may therefore impede the formulation studies. Continuous trends for the log-linear and spline regression can be observed, just as for the neural network and therefore these models will be included for further accuracy assessment.

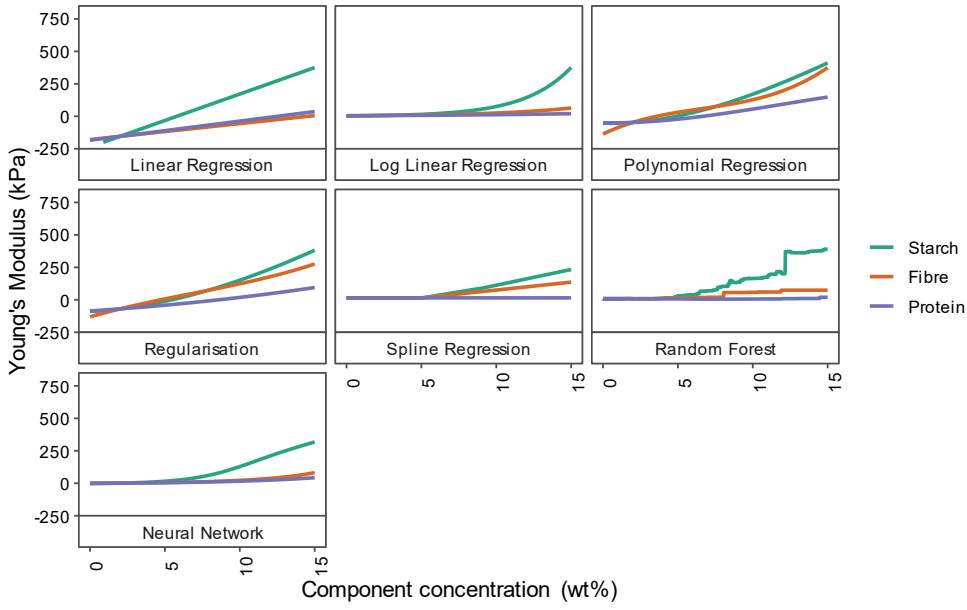

Figure 5.5 Feasibility of behaviour of gel stiffness models based on main macro components with a varying component concentration between 1-15 wt% with the other components fixed at 2 wt%.

5.3.3 Model metrics for gel stiffness

The next step is to evaluate the model metrics of the considered models, with a focus on the log-linear regression, spline regression, and neural network (main macro components in Table 5.2 and the others in the Appendix). In general, all linear regressions and the regularisation overfit the test data less compared to the other

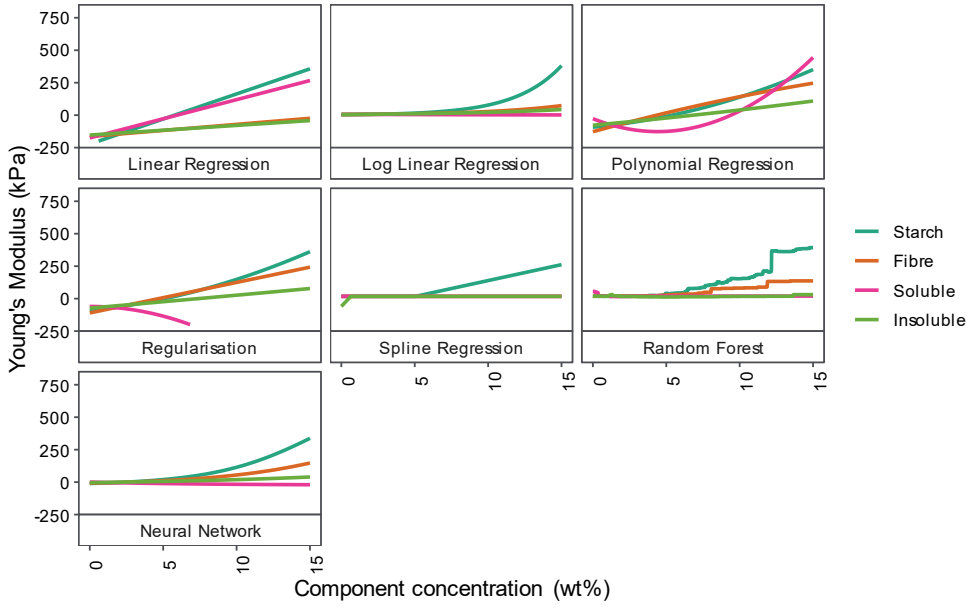

Figure 5.6 Feasibility of the behaviour of gel stiffness models based on main macro components with protein split according to solubility with a varying component concentration between 1-15 wt% with the other components fixed at 2 wt%.

models, since the RMSE values of the train and test set are relatively close to each other.

The log-linear regression model has a relatively high train and test RMSE due to the calculation of the model metrics with exponentiated (back-transformed) Young's modulus. To get more accurate ingredient formulations, the focus will be on the other two models. For the training set, the model metrics of the spline regression and neural network are similar. The test errors of the neural network are smaller than the spline regression (RMSE), the predicted and observed values have a higher correlation (R^2), and residuals are more evenly distributed (MAE). This shows that the neural network is more stable in predicting new samples based on this dataset. Nevertheless, the test errors are a lot higher compared to the training set, which indicates that some overfitting takes place. This can mostly be attributed to the relatively small dataset for a neural network. There are ways to reduce overfitting and are well summarized in literature[32,172,173] but out of scope for this study. In

addition, the test errors are still in an acceptable range considering the range of Young's Modulus of the measured samples.

An additional requirement of the approach is that the models are suitable for interpolation (Table 5.3). The neural network only scores better in the test set performed at 15 wt% and at 21 wt% the spline regression performs slightly better. Nevertheless, based on all model metrics, the stiffness of gels in this study is most accurately described by a neural network using the main macro components. The next step is to evaluate if a split in protein according to processing history improved the prediction accuracy.

To compare the model metrics of the different sets of variables it is important to consider the variance in the metrics caused by the k-fold cross-validation[173] and random starting weight values of each node in the neural networks[168]. Therefore, conclusions on the most suitable set of variables are based on the significant differences between the models that have been repeated five times (Figure 5.7). We conclude that there is no significant difference (95% CI, $p<0.05$) between the sets of variables. This contrasts with the expected effect of protein solubility and nativity. However, starch is the dominant component in the mixtures and may therefore outweigh the contribution of soluble and insoluble protein (or native and denatured).

To summarize, all model types were evaluated for three sets of independent variables: the main macro components and the main macro components with a protein split according to the nativity as well as the solubility. The neural network with the main macro components trained on the dataset in this study showed feasible predictions for values near boundaries, exhibits plausible physical behaviour, and have acceptable model metrics. Splitting protein according to processing history did not significantly improve the model predictions, as can also be seen in the parity of the selected model in Figure 5.8. Therefore, we conclude that a neural network with the main macro components as input variables is most suitable to predict this dataset of gel stiffness.

Table 5.2 Model metrics (MAE, RMSE, R^2) models for main macro components analysis.

Model	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Linear Model	49.51	0.93	38.95	71.30	0.83	48.46
Log Linear Model	216.89	0.77	82.55	169.81	0.73	86.16
Poly Model	48.30	0.93	36.94	47.43	0.92	35.14
Regularisation Model	49.48	0.93	37.47	49.22	0.91	34.89
Spline Model	29.97	0.97	20.93	74.87	0.82	46.52
Random Forest	25.09	0.98	16.65	85.63	0.74	55.46
Neural network	28.91	0.98	19.82	68.50	0.87	44.85

Table 5.3 Model metrics (MAE, RMSE, R^2) models interpolated at 15 and 21 wt% for main macro component analysis.

Model	RMSE 15%	R^2 15%	MAE 15%	RMSE 21%	R^2 21%	MAE 21%
Linear Model	38.76	0.90	29.27	38.26	0.94	34.06
Log Linear Model	48.73	0.83	36.67	69.02	0.88	53.36
Poly Model	34.69	0.79	27.46	42.19	0.93	36.98
Regularisation Model	35.16	0.82	28.29	42.40	0.93	36.99
Spline Model	28.47	0.88	21.95	58.94	0.95	50.91
Random Forest	66.65	0.72	46.21	50.95	0.91	42.51
Neural Network	29.46	0.85	20.77	65.92	0.96	57.50

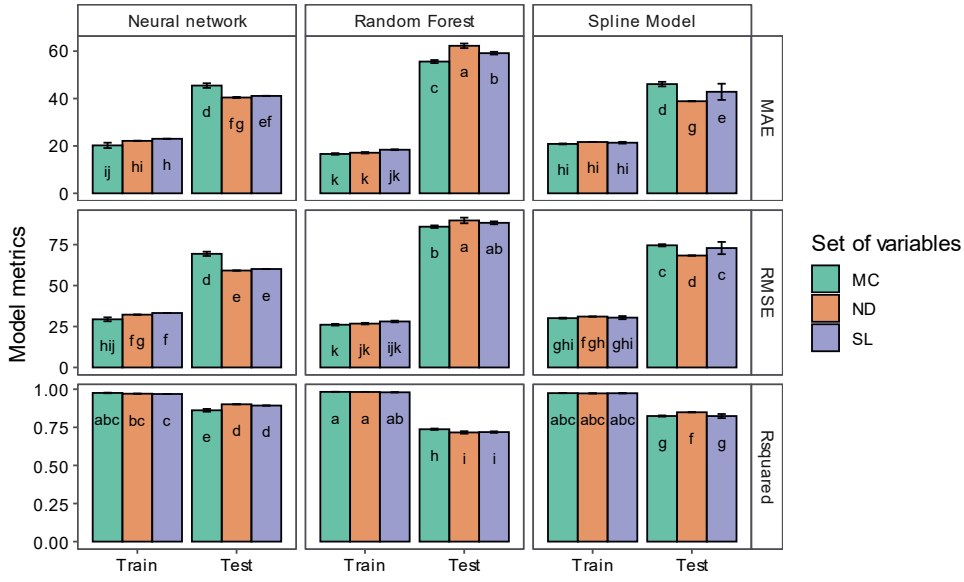

Figure 5.7 Comparison model metrics for gel stiffness of the spline regression, random forest, and neural network run 5 times to obtain the mean and standard deviation using three sets of variables: main macro components (MC) and main macro components with a split in protein according to nativity (ND) and solubility (SL). The letters indicate the significant difference within each metric (95% CI, $p \leq 0.05$.).

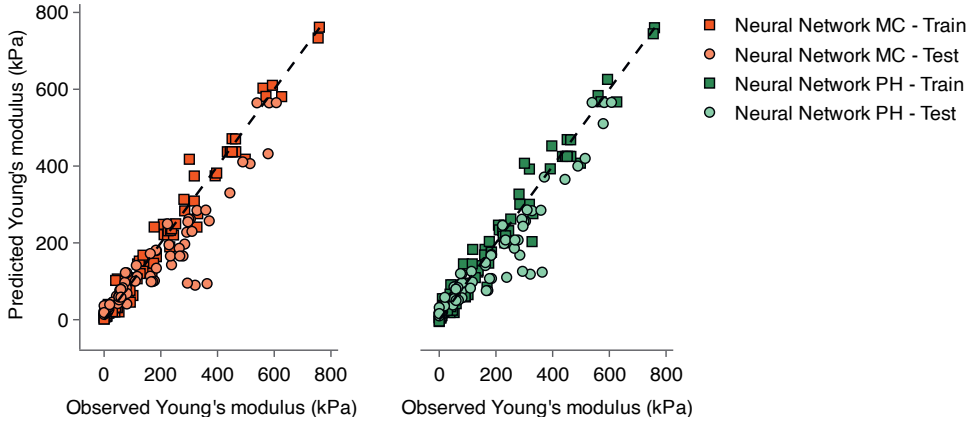

Figure 5.8 Parity of the training sets for two final selected models for the main macro components (MC) and effect of processing history (PH) with regards to a split in protein solubility.

5.3.4 Revisiting the quantification of the final viscosity

Previously we described the final viscosity of the same ingredients as used in this study using linear regression on a log scale with the main macro components, as a split according to nativity was found to have limited effect.[129] Even though this regression was sufficient for describing the final viscosity, it is of interest to find an alternative model without transformation bias. Therefore, the approach proposed in this study is applied to the viscosity dataset as well. All different sets of variables and models were evaluated and showed no significant differences (Figure A 5.6) and therefore only the results from the main macro components are shown below.

The feasibility of the predictions close to boundary constraints in regard to viscosity mostly concerns dispersions with a very low viscosity. The linear and polynomial regression, and regularisation predict many and large negative values (Figure 5.9A) and are therefore excluded from further analyses. Furthermore, the behaviour of the random forest levels off at 8% starch, which is rather unphysical due to the contribution of starch to the final viscosity (Figure 5.9B). The neural network predicts some negative values, which can be neglected due to their magnitude. The spline regression, and also most other model types, estimates the behaviour of fibre very low, with a negative trend after approximately 5 wt%, which is outside of the measuring range and therefore acceptable.

Table 5.4 summarizes the metrics of all model types with the main macro components as the set of variables. The spline regression shows the highest potential and has acceptable test errors. The comparison of the spline regression to the linear regression on the log scale shows that spline regression is a better alternative to the log-linear model presented in our previous study (Figure 5.10). Notably, in the previous study, it was concluded that high fibre samples were not predicted well using the log linear model, which was later suggested to be caused by phase separation taking place at these fibre concentrations.[174] Based on the predicted viscosity values, it seems that the spline regression is able to describe this change in behaviour. This can be attributed to the non-parametric character of the spline regression, which makes it more flexible to fit data as opposed to the predefined linear regression.

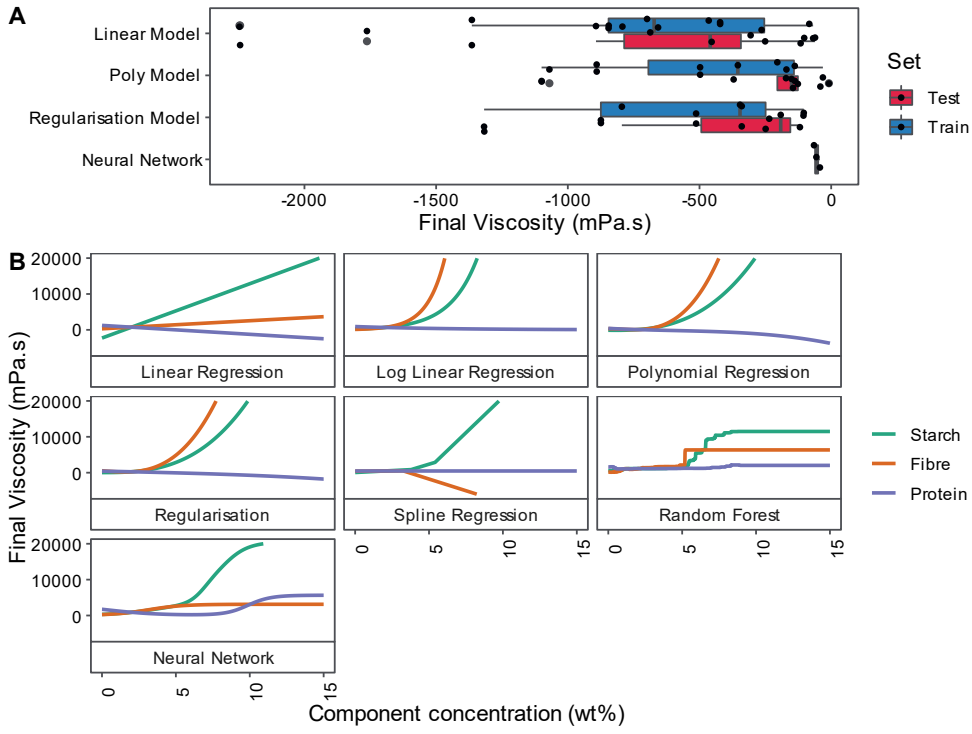

Figure 5.9 Feasibility of thickened dispersions at the lower boundary with model based on main macro components. A) Distribution of the data points for which model fits (training set in blue) or predicts (test set in red) a negative B) Feasibility of behaviour of final viscosity models based on main macro components, with a varying component concentration between 1-15 wt% with the other components fixed at 2 wt%.

Table 5.4 Model metrics (MAE, RMSE, R^2) models predicting the final viscosity data.

Model	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Linear Model	1217.57	0.84	906.52	1371.44	0.72	1200.10
Log Linear Model	1847.20	0.81	806.04	1503.29	0.95	627.29
Poly Model	764.60	0.94	472.39	696.80	0.91	482.79
Regularisation Model	772.01	0.94	480.65	646.87	0.91	448.98
Spline Model	420.60	0.98	234.30	597.53	0.95	457.04
Random Forest	501.39	0.98	242.15	887.85	0.94	682.24
Neural network	411.68	0.98	215.28	730.16	0.99	440.95

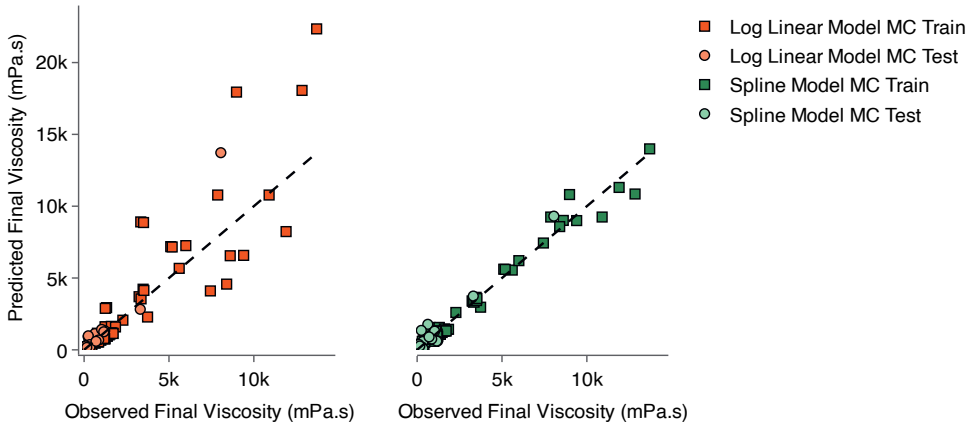

Figure 5.10 Parity plots for the prediction of the final viscosity of main macro components of log linear model and spline regression.

5.4 Conclusion

We proposed a procedure to systematically select suitable machine learning models for quantifying functional properties of (complex) food ingredients as a function of variables describing the compositions and processing history. This procedure was evaluated for the stiffness of yellow pea gels. Next to correctly predicting the values from the training and test set, other requirements of the model were to not return any physically impossible values (such as negative values or unrealistic inflection points) and to correctly identify compositions that do not gel.

It was not possible to avoid negative values for gels with low stiffness when using linear regression, except on the log scale. This was not solved by regularisation of the linear polynomial regression. The random forest showed discontinuous behaviour in all cases, which could conflict with formulation optimisation and was therefore not preferred. The linear regression on a log scale, the spline regression, and the neural network all performed well in the feasibility studies, as these model types did not result in negative values and showed feasible physical behaviour. Based on the model metrics of the complete dataset as well as of the interpolation test sets, we decide that a neural network with the main macro components as independent variables is most suitable to describe the gel stiffness of yellow pea ingredients used in this specific study.

To further assess the systematic framework, the same evaluation was performed with an existing dataset for viscosity. While in the original paper a log scale was found suitable, here a spline regression emerges as an alternative model that is not sensitive to transformation bias. This shows that it remains important to evaluate a range of model types to describe techno-functional properties and to select a model that has both good model metrics and shows physically feasible behaviour.

This study on gelation shows that it is possible to use machine learning to describe properties that depend non-linearly on the independent compositional variables. This is important for the incorporation of new, complex ingredients into high-quality foods. Selected models can be used to identify a formulation window that ensures the right functional properties of the final food even when using complex, mildly processed ingredients. In this window, one can then optimize for the smallest possible environmental impact.

5.5 Acknowledgements

Many thanks to Zeger Kievit for his contributions to developing the protocol to measure gel stiffness.

5.6 Appendix

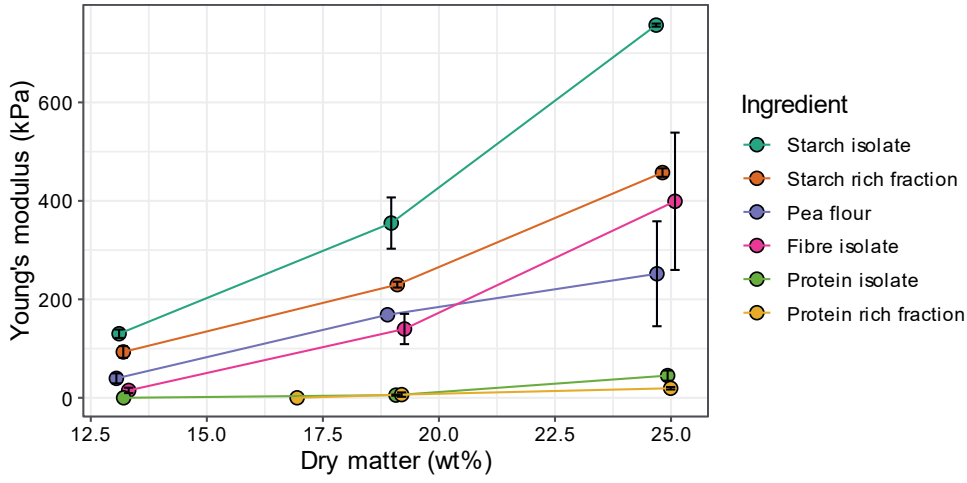

Figure A 5.1 Gel stiffness of individual ingredients at different concentrations.

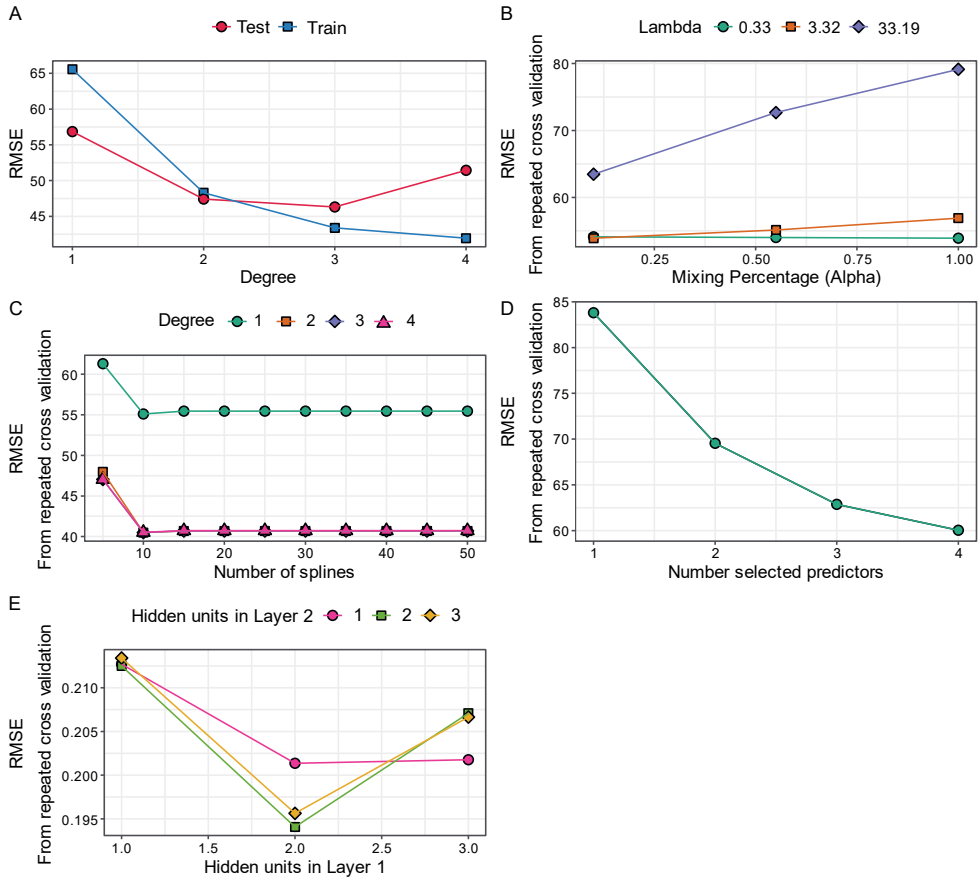

Figure A 5.2 Hyperparameter tuning for gel stiffness models based on main macro components: A) the degree of polynomial regression, B) alpha and lambda determination for regularisation (package glmnet version 4.1.)[175], C) the degree and number of spline determination for spline regression (package earth version 5.3.1)[176], D) the number of predictors used for the random forest (package randomForestversion 4.7.1.) [177], and lastly E) the number of layers and nodes for the neural network (package neuralnet version 1.44.2)[178]. The degree of the polynomial model is based on the RMSE value of the test and training set, whereas the other models are based on the RMSE from cross-validation.

Table A 5.1 Tuned hyperparameters for all models to predict gel stiffness.

Functional Property	Variable Set	Degree Polynomial (n)	Regularisation (alpha, lambda)	Splines (nr, degree)	Random Forest Predictors (nr)	Hidden neural nodes (layer 1, 2, 3)
Gel	Main macro components	3	0.1, 3.32	10, 3	4	2, 2, 0
Gel	Nativity split	2	0.1, 0.33	10, 3	4	1, 2, 0
Gel	Solubility split	2	0.55, 3.32	10, 4	4	1, 2, 0
Viscosity	Main macro components	3	0.55, 5.17	15, 2	3	3, 2, 0

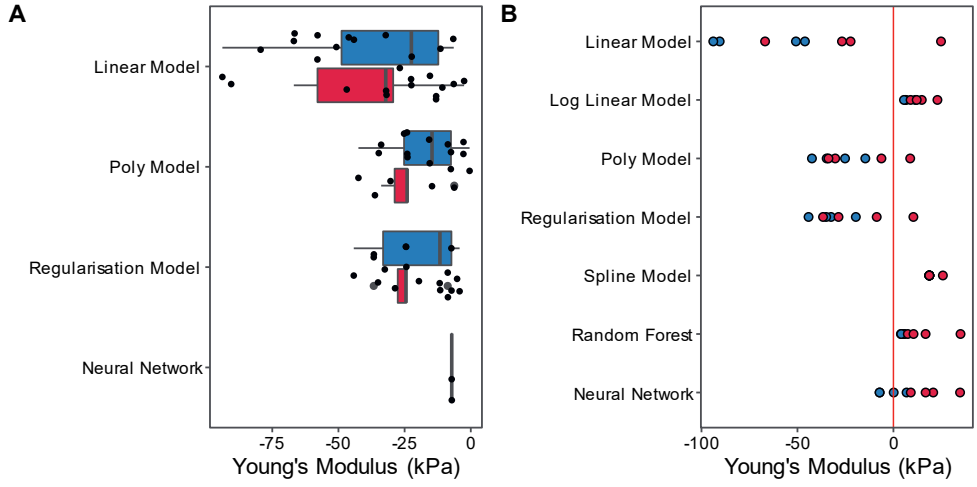

Figure A 5.3 Feasibility gels at lower boundary with model based on main macro components with protein split according to nativity. A) Distribution of the data points for which model fits (training set in blue) or predicts (test set in red) a negative B) model fits (training in blue) or prediction (test in red) for gels that did not gel (0 kPa). Please note that in B some samples are overlaying due to a similar prediction, but for all models the same number of samples is plotted.

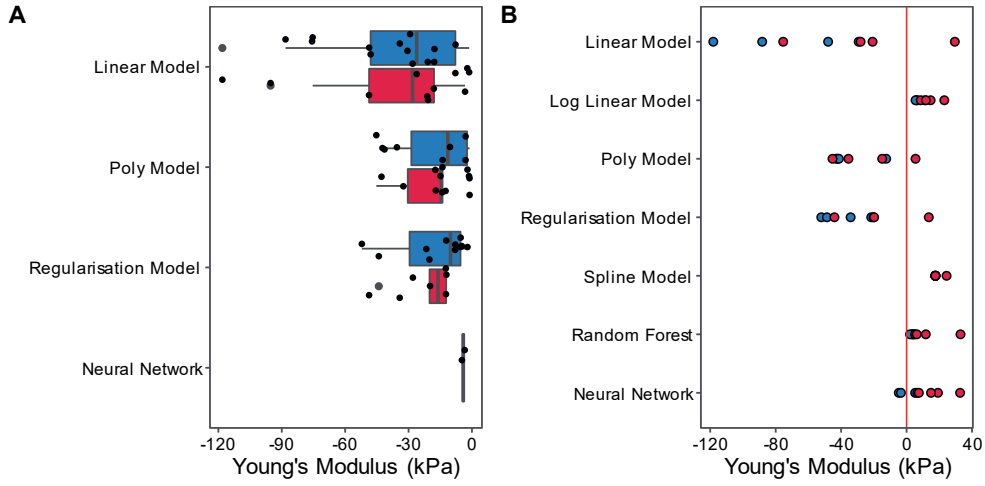

Figure A 5.4 Feasibility of gels at the lower boundary with model based on main macro components with protein split according to solubility A) Distribution of the data points for which model fits (training set in blue) or predicts (test set in red) a negative B) model fits (training in blue) or prediction (test in red) for gels that did not gel (0 kPa). Please note that in B some samples are overlaying due to a similar prediction, but for all models the same number of samples is plotted.

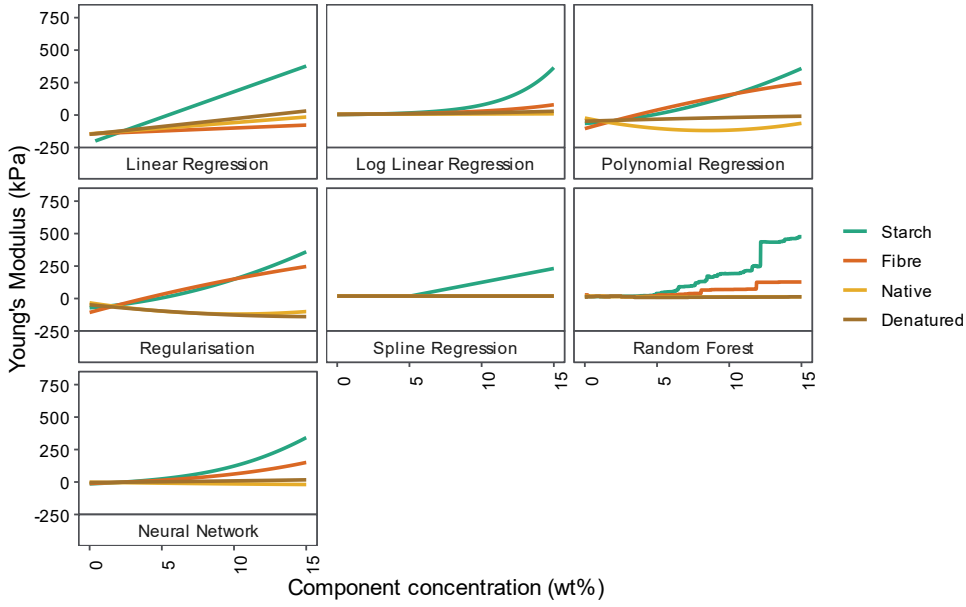

Figure A 5.5 Feasibility of behaviour of gel stiffness models based on main macro components with effect processing history, with a varying component concentration between 1-15 wt% with the other components fixed at 2 wt%.

Table A 5.2 Model metrics (MAE, RMSE, R^2) models for gel stiffness models with main macro components and the protein split according to nativity.

Model	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Linear Model	48.08	0.93	37.17	48.45	0.91	40.58
Log Linear Model	192.77	0.81	81.18	158.54	0.73	81.12
Poly Model	45.19	0.94	34.32	44.27	0.93	33.74
Regularisation Model	45.34	0.94	34.29	44.42	0.93	33.62
Spline Model	32.57	0.97	23.01	84.25	0.78	53.31
Random Forest	26.01	0.98	16.18	87.42	0.73	60.79
Neural network	32.32	0.97	22.21	59.28	0.90	40.58

Table A 5.3 Model metrics (MAE, RMSE, R^2) models for gel stiffness interpolated at 15 and 21 wt% with main macro components and the protein split according to nativity.

Model	RMSE 15%	R^2 15%	MAE 15%	RMSE 21%	R^2 21%	MAE 21%
Linear Model	38.64	0.90	32.61	36.73	0.95	33.36
Log Linear Model	47.97	0.83	36.58	73.95	0.87	59.78
Poly Model	28.08	0.87	17.38	45.68	0.91	39.31
Regularisation Model	28.71	0.86	18.96	44.41	0.92	38.44
Spline Model	30.17	0.84	22.65	67.37	0.92	56.40
Random Forest	60.49	0.78	48.05	55.73	0.89	42.59
Neural Network	30.15	0.91	24.57	63.44	0.93	52.40

Table A 5.4 Model metrics (MAE, RMSE, R^2) models for gel stiffness models with the effect of solubility with main macro components.

Model	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Linear Model	47.72	0.94	36.49	69.51	0.83	48.08
Log Linear Model	167.87	0.84	74.63	153.99	0.74	79.80
Poly Model	46.15	0.94	34.91	45.93	0.92	34.55
Regularisation Model	48.60	0.93	36.09	50.14	0.91	36.62
Spline Model	33.64	0.97	23.96	81.98	0.78	52.79
Random Forest	28.59	0.98	18.80	88.98	0.71	60.12
Neural network	33.15	0.97	23.04	60.24	0.89	40.77

Table A 5.5 Model metrics (MAE, RMSE, R^2) models for gel stiffness interpolated at 15 and 21 wt% with main macro component analysis with the effect of solubility.

Model	RMSE 15%	R^2 15%	MAE 15%	RMSE 21%	R^2 21%	MAE 21%
Linear Model	34.90	0.90	30.30	48.58	0.91	42.08
Log Linear Model	47.73	0.85	36.06	66.21	0.89	50.04
Poly Model	29.88	0.85	19.47	47.21	0.91	41.43
Regularisation Model	34.49	0.84	27.79	47.23	0.91	40.36
Spline Model	28.67	0.85	24.43	63.34	0.92	54.78
Random Forest	74.08	0.66	54.37	47.48	0.92	39.23
Neural Network	30.29	0.90	24.40	64.89	0.92	53.84

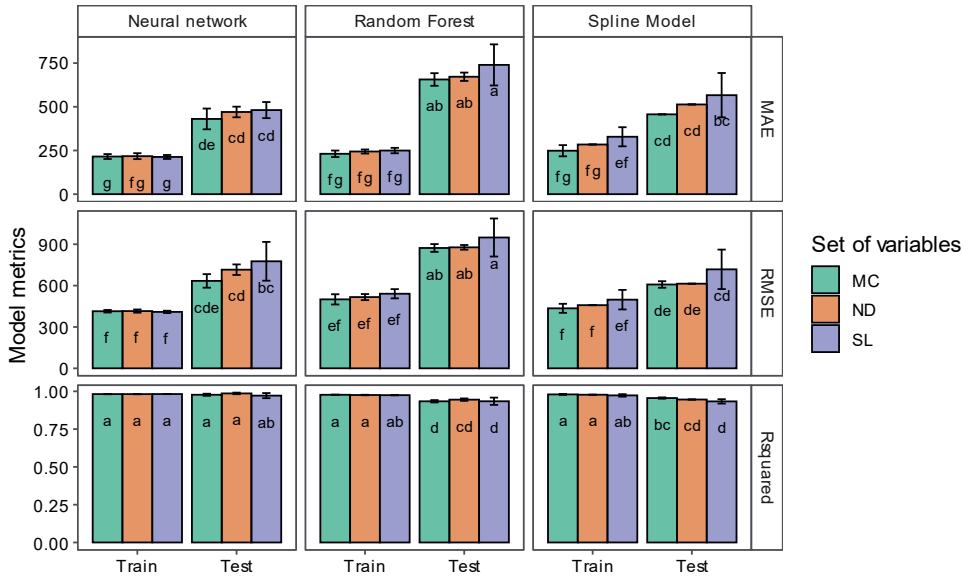

Figure A 5.6 Comparison model metrics (MAE, RMSE, R^2) for the final viscosity of the spline regression, random forest, and neural network run 5 times to obtain the mean and standard deviation using three sets of variables: main macro components (MC) and main macro components with a split in protein according to nativity (ND) and solubility (SL). The letters indicate the significant difference within each metric (95% CI, $p \leq 0.05$.

Chapter 6

Mixed crops

Quantifying techno-functional properties of ingredients from multiple crops using machine learning

---

A. Lie-Piang, J. Hageman, I. Vreenegoor, K. van der Kolk, S. de Leeuw, A. van der Padt, R.M. Boom

Submitted

Abstract

Mildly refined ingredients from plants consist of multiple components. Therefore, they benefit from a food formulation based on techno-functional properties rather than composition. The techno-functional properties of these complex mixtures can be quantified using machine learning. To increase the space in which ingredients can be formulated, it is of interest to also quantify these properties of ingredients from multiple crops. In this study we quantify and predict the techno-functional properties of ingredients from yellow pea and lupine seeds combined and compared them to models based on single crops. The properties quantified were gelation, viscosity, emulsion stability, and foaming capacity. The relationships were quantified using spline regression, random forest, and neural network. Suitable models were picked based on model accuracy and physical feasibility of model predictions. A single model could be created for each techno-functional property, albeit with a trade-off of higher prediction errors as compared to models based on individual crops. A reflection on sizes of the datasets showed that they could be reduced for some properties. The approach presented in this study contributes to the adoption of milder refined ingredients to enhance the sustainability in food production.

6.1 Introduction

Food ingredients with a lower extent of refining are more resource use efficient [8,9], which in turn reduces the environmental footprint of food products. Less refined ingredients are alternative ingredients that may replace conventional isolates, such as protein or starch, that are commonly used to texturize food products. Air classification, electrostatic separation, and mild aqueous fractionation are all examples of methods that deliver functional ingredients, rather than pure conventional isolates.[13–15] As these fractions are not pure, it was proposed earlier to blend the ingredients to match the final composition of food products.[18] Unconventional ingredients could not always be matched due to their complex compositions. This may not be necessary if a different composition could result in a product with the same functional properties. Therefore, food formulation practices would benefit from ingredient selection based on their functional properties, instead of composition.

Mathematical modelling or machine learning can be employed to quantify functional properties and subsequently formulate products, as was shown for example for chocolate cookies.[21] Ingredients can also be selected for formulations based on their structural contribution to food products, which was shown for the thickening capacity of yellow pea ingredients.[129] In this study, formulations with different compositions and environmental impacts could be generated using the functionality-driven formulation approach using least square regression with the composition of the ingredients as independent variables. To quantify other functional properties that are not linear, a framework was proposed to select a suitable machine learning algorithm.[179] The models were subjected to the condition that the predictions were physically feasible, such as the absence of artefacts that could impede with the formulation. Since it could not be expected that one single algorithm always results in an acceptable quantification, a range of methods was explored.

While the previous work did demonstrate the principle of functionality-driven formulation, it was based on just one raw material. In practice, generally, more than one raw material is used to create a set of ingredients with suitable properties.

Therefore, the proof of principle of the functionality-driven formulation should be extended to include multiple crops rather than one. Frequently used crops for ingredients are for example grains or legumes. Although most crops consist of the same basic building blocks protein, starch, fibre, oil, and sugars, they can yield different techno-functional properties. For example, fibres from different sources have different solubilities and water holding capacities.[180] This complicates the optimization of ingredient blends made with different crops. Functionality-driven food product formulation would benefit from models that can quantitatively predict a functional property from multiple crops and hence have more general validity.

This study therefore assesses the potential to quantify and predict a techno-functional property of ingredients from multiple crops using a single model. The properties that will be assessed are gelation, viscosity, emulsion stability, and foaming capacity. This will be exemplified with conventional, high-purity ingredients and with mildly refined ingredients from yellow peas and from lupine seeds. First, the relation between the techno-functional properties and the composition of the ingredients will be discussed. Next, the relation between the composition of the ingredients and each techno-functional property will be quantified for each crop individually using machine learning algorithms. To assess whether models for multiple crops can be created for each property, new single models will be fitted to the data from all crops. These will be compared to the models based on the individual crops. We conclude with a critical discussion on the accuracy of the results as well as a reflection on the size of the dataset that is required to quantify each techno-functional property.

6.2 Materials & methods

6.2.1 Ingredients

Pre-dried yellow peas (*Pisum sativum L.*) were purchased from Alimex (The Netherlands). Pre-dried lupine seeds (*Lupinus angustifolius*) were obtained from InvejaFood (The Netherlands). Both yellow pea and lupine seed ingredients were produced using a pin mill (LV 15M, Condux-Werk, Germany), ZPS50 impact mill (Germany), and an ATP50 classifier (Hosokawa-Alpine, Germany) in a room with a controlled relative humidity of 30% and room temperature. A thermometer inside

the mill indicated a temperature between 16 and 34°C. The water used in all experiments was deionized in a Milli-Q purification system (Merck Millipore, Burlington, USA). Sunflower oil for the emulsions was purchased from a local supermarket.

6.2.1.1 Commercial isolates

Commercial isolates from yellow pea were a pea protein isolate (YPI) (Nutralys F85G) and a pea fibre isolate (YFI) (PEA FIBRE I 50 M), purchased from Roquette Frères S.A (St. Louis, USA) and a starch isolate from Emsland Stärke GmbH (Germany). A commercial protein isolate from lupine seeds (LPI) was obtained from Prolupin GmbH (Germany) and a fibre isolate slurry as well. The fibre isolate slurry was freeze dried in a Pilot freeze dryer (Christ Epsilon 2-6D, Germany). The dried fibre isolate was subsequently milled in an impact mill with a mill speed set at 8000 rpm and an airflow of 75 m^3/h and a classifier wheel speed of 2200 rpm. The feed rate was set at 7 rpm.

6.2.1.2 Mildly refined ingredients

The hulls were first removed from the peas with a Satake TM05 pearling machine (Satake Corporation, Japan). A yellow pea flour (YPF) was then produced by pre-milling the peas into grits using a pin mill at room temperature and then milling using an impact mill with an impact mill speed of 8000 rpm, airflow at 52 m^3/h, and classifier wheel speed at 4000 pm, and a feed rate of 2 rpm (method adopted[49]).The flour was subsequently air classified into a fine, protein-enriched fraction (YFF) and a coarse, starch enriched (YCF) fraction with a fixed airflow at 52 m^3/h, classifier wheel speed at 5000 rpm, and feed rate at 20 rpm.

Lupine seeds were used to prepare lupine flour (LF) from the non-dehulled seeds. First, the seeds were pre-milled with the pin mill, after which they were milled with a feed rate of 7 rpm in the impact mill at a milling speed of 8000 rpm and classifier wheel speed of 2200 rpm and an airflow of 75 m^3/h. A fine protein-rich fraction (LFF) and a coarse fibre-rich fraction (LCF) were subsequently produced with an air classifier at a feed rate of 20 rpm, classifier speed of 10,000 rpm, and an airflow of 80 m^3/h (method adapted[48]). For all in-house produced lupine ingredients, larger

pieces of the hull were removed using a 0.16 mm grid sieve (Retsch, Germany) and later with an air jet sieve (Hosokawa-Alpine, Germany) with a grid size of 0.4 mm.

6.2.1.3 Compositional analysis

The Kjeldahl method with a nitrogen conversion factor of 5.52[137] for pea protein was used to determine the protein content. The starch concentration in the samples was determined with a Total Starch Amyloglucosidase – α-Amylase Assay Kit, AOAC Method 996.11 (Megazyme International Ireland Ltd, Bray, Ireland). Next, AOAC method 985.29 was used to determine the total dietary fibre. The solids were determined by drying the ingredients in a vacuum oven at 70 ˚C. The rest fraction was determined as the difference between the dry matter and the protein, starch, and fibre content and contains among others oil, ash, salts, and sugars Table 6.1.

Table 6.1 Composition of ingredients based on g/100 g dry matter and nitrogen solubility index (NSI%) with mean and standard deviation. Please note that there is a yellow pea flour only used for the viscosity measurements, due to different batches.

	Protein	+/-	Starch	+/-	Fibre	+/-	Rest	+/-	Solids	+/-	NSI%	+/-
YPI	74.1	1.8	0.4	0.0	2.2	1.2	23.3	1.6	93.7	0.5	16.4	5.8
YSI	0.5	0.0	91.4	2.9	0.8	0.8	7.3	0.0	88.4	0.5	0.0	0.0
YFI	7.8	0.5	33.8	0.5	53.6	4.6	4.8	1.1	95.2	0.5	34.7	8.9
YFF	44.4	1.1	6.9	0.2	18.9	2.9	29.8	2.1	93.5	0.5	48.0	0.8
YCF	10.5	0.5	66.6	1.9	4.6	1.5	18.3	1.6	90.9	0.5	55.3	7.5
YPF*	22.3	0.8	47.9	7.1	10.6	0.1	19.2	0.4	90.6	0.5	61.3	0.9
YPF	25.2	0.7	48.2	0.5	9.3	2.1	17.4	0.0	90.6	0.5	61.3	0.9
LPI	92.2	2.3	0.2	0.0	0.2	0.0	7.5	0.7	97.7	0.5	60.1	0.9
LFI	16.3	0.6	1.2	0.0	79.6	5.7	3.2	0.5	93.5	0.5	47.5	5.4
LF	36.8	1.0	0.2	0.0	39.3	4.1	23.7	2.2	92.8	0.5	21.7	1.5
LFF	55.1	1.4	0.2	0.0	14.6	2.5	30.1	2.4	94.4	0.5	19.4	0.5
LCF	30.4	0.9	0.2	0.0	49.0	4.6	20.4	2.0	92.2	0.5	33.3	3.7

*yellow pea flour batch used for the viscosity measurements

The solubility of the protein in each ingredient was expressed with the nitrogen solubility index (NSI%) and was determined using an adopted protocol.[179] Each ingredient was mixed thoroughly in a 1 wt% dispersion and rotated for 30 minutes, after which they were centrifuged (30 min, 10,000 g, at 20 ˚C). The pellet and

supernatant were carefully separated. The pellet was further weighted in the tube and subsequently dried in an oven at 105 °C for 24 hours. From the dried pellet the dry matter and protein content were determined. With the latter, the soluble protein was expressed in the NSI% as a percentage of soluble protein of the total protein present in the ingredients. All protein in yellow pea is considered denatured in the isolates[10] and native in the mildly refined fractions. As the distinction in nativity between isolates and mildly refined ingredients in lupine ingredients was not that clear, this was not considered in the study (data not shown).

6.2.2 Selection of ingredient blends

The models in this study should be able to predict the techno-functional properties of all possible ingredient blends. The ingredient blends were chosen in such a way that the space in which ingredients can be formulated was covered as best as possible (high and low starch, fibre, protein and the rest of the components) as described in another study.[129] All yellow pea (YPI, YSI, YFI, YFF, YCF, and YPF) and lupine seed ingredients (LPI, LFI, LFF, LCF, LF) were therefore blended in the following combinations:

- yellow pea mixtures (isolates and mildly refined individually and mixed)
- lupine mixtures (isolates and mildly refined individually and mixed)
- yellow pea and lupine mixtures (isolates and mildly refined individually and mixed)

Figure 6.1 depicts the ratio between protein, starch, and fibre for the three categories described above (with rest fraction in Appendix Figure A 6.1). The rest fraction represents all dry components besides protein, starch, and fibre. All ingredient blends depicted in these Figures are measured between 1-25 wt%, which is discussed in section 6.2.4.1.2 in more detail.

6.2.3 Techno-functional properties

The emulsion separation velocity, foamability, gel stiffness, and final viscosity were measured for all selected ingredient blends. Dry powders were always premixed and dispersed in deionized water at room temperature. All ingredient blends were

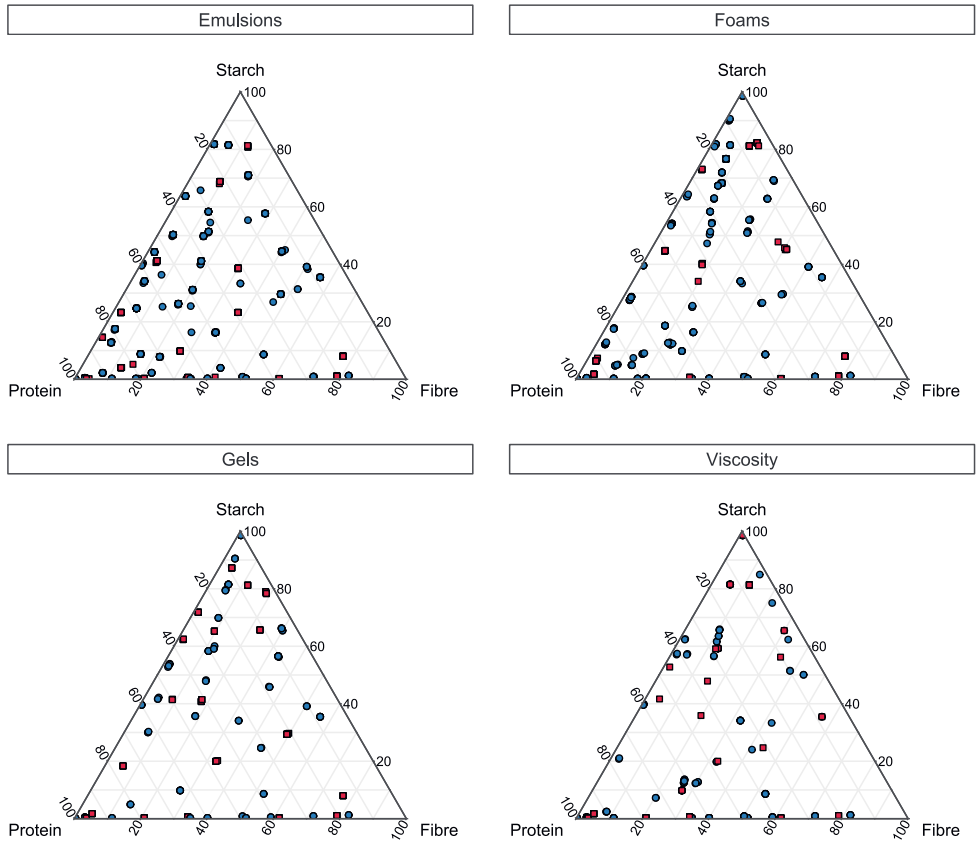

Figure 6.1 Ternary plots of the ingredient blends selected in this study, presenting their composition (ratio protein, starch, and fibre) for measuring the 4 techno-functional properties train (●) and test (■) observations. All determined in a range of 1-25 wt%.

measured once, with the exception of some duplicates or triplicates for protocol development.

6.2.3.1 Emulsion separation velocity

The emulsion separation velocity (μm/s) was measured for 10 wt% (sunflower) oil emulsions with a total weight of 150 g. The aqueous phase was a dispersion of the selected ingredient ratios, which were hydrated for 30 minutes under gentle agitation while covered with parafilm. The emulsion preparation was adapted from another study. [181] The coarse emulsion was made using a rotor-stator homogenizer

(Ultraturrax IKA T18 basic, Germany) at 11,000 rpm for 1 minute. This pre-emulsion was further homogenized in a colloid mill (IKA Magic Lab, Germany) with a gap width of 0.16 for two minutes at 15,000 rpm. The separation (or creaming) velocity was measured with an adapted protocol with a LUMiFuge (LUM LUMGmbH, Germany) at constant gravitational acceleration of 2,227 rpm at room temperature with a light factor of 1.0.[182] Measurements were taken every 15 sec with 240 measuring points in total, which is equivalent to a storage time of approximately 30 days at 1g. The separation velocity (μm/s) was calculated using the LUMiFuge Front Tracking module at a transmission of 25% with equation 6-1 in which ΔL is the change of the position of the layer measured at 25% transmission for a time period Δt, only in the linear part.

$$separation\ velocity = \frac{|\Delta L|}{\Delta t} \tag{6-1}$$

Since the LUMiFuge system measures the change of a creaming layer over time, we cannot exclude any effect of sedimentation on the position of the layer at 25%. Sedimentation can be observed by a movement downwards instead of upwards and measurements that clearly showed this behaviour were excluded. Dispersions that were too thick to measure were considered 'stable' in this method, i.e., having a separation velocity of 0 μm/s.

6.2.3.2 Foamability

The foamability, or overrun (%), was determined using an adapted protocol.[93] The dry powders were dispersed using gentle agitation with a magnetic stirrer for 30 minutes. The samples were always covered with parafilm. 10 mL of the dispersion was subsequently foamed in a plastic tube (diameter 33.5 mm) using an overhead stirrer for 2 min minutes at 2000 rpm with a froth (Aerolatte, UK). The foaming tubes were fixated at an angle of 45° to facilitate the incorporation of air and increase the reproducibility of the measurements. The foam height was directly captured by a camera and was recalculated to the volume. The foam overrun was determined with equation 6-2.

$$Foam\ overrun\ (\%) = \frac{Foam\ volume\ (mL)}{Liquid\ volume\ (mL)} * 100 \tag{6-2}$$

6.2.3.3 Gel stiffness

The gel stiffness in terms of Young's modulus was measured according to an existing protocol.[179] The dispersions were initially hydrated for 5-10 minutes and then hydrated for 1 hour while stirring gently and covered with parafilm. The dispersions were subsequently transferred to Teflon tubes (20 mm diameter and 100 mm length) and heated in a rotating water bath at 90 °C for 30 minutes at a speed of 30 rpm. After heating the tubes were removed from the water bath and left to rotate at room temperature for one hour to cool down and prevent sedimentation. The tubes were further cooled overnight in the fridge at 4 °C and then again removed from the fridge one hour before measurement. The gels were released from the tubes and cut into 3 mm wide and 20 mm long cylinders. The Henky's stress and strain[161] were measured and calculated for all three cylinders and averaged.

6.2.3.4 Final viscosity

The heated final viscosity was measured according to the method described in another study,[129] in which the dispersions were first hydrated for 5 – 10 minutes, depending on the consistency. The viscosity was determined with an Anton Paar Rheometer MCR301 equipped with a starch cell C-ETD160/ST (Anton Paar, Austria). The dispersions were stirred at 600 rpm for 10 s at 50 °C to reassure a thorough dispersion prior to starting the measurement. The viscosity measurements were performed at a constant rotational speed of 160 rpm, and a temperature–time profile was used to heat all samples to 95 °C and subsequently cool them to 50 °C in a total of 25 min. Measurement of the unheated final viscosity was performed according to the same protocol without the heating ramp.

6.2.4 Data analysis

6.2.4.1 Quantification of techno-functional properties

6.2.4.1.1 Model training, testing, and selection

The relations between the selected ingredient blends and the techno-functional properties were quantified using a range of machine learning methods according to the framework presented in a previous study.[179] For a detailed description of each method we refer to this article. The following statistical and machine learning

algorithms were used: 1) a stepwise AIC multiple linear regression, 2) the same stepwise AIC linear regression but with a log transformed output, 3) polynomial linear regression (power 2-4), 4) a regularized form of the polynomial linear regression to reduce parameters, 5) a spline regression, (6) a random forest, and (7) a neural network. For the neural network, the dataset was centred and scaled and fitted with a Sigmoid function. All models were fitted using adopted stepwise linear regression or the Caret package[162] in RStudio V4.1.0 with the code published on Git@WUR.[183]

All models were trained and tested using independent datasets, which is explained in more detail below. The regularized polynomial linear regression, spline regression, random forest, and neural network models contain hyperparameters, which set the architecture of the models (e.g., number of hidden nodes in the neural network). These hyperparameters were chosen based on the average error obtained from 5 times repeated 10-fold cross validation from only the training set. Next, all models were tested again on the separate test dataset.

The foam dataset was initially filtered for dispersions that had a solid structure since the foam overrun could not be measured precisely. These solid and liquid foams were distinguished using a simple decision tree [184] with the unheated viscosity as the input variable. The accuracy and resulting confusion matrices were provided to compare the performance of the classification, with the addition of the kappa value. The kappa value provides a correction for the random chance of values ending up as either liquid or solid foams.[185] Only the liquid foams were extracted from the dataset and subsequently used in further modelling steps.

For each functional property, a final model was selected based on three conditions: the selected model 1) does not return unfeasible negative values 2) shows physically feasible behaviour of each component without any artefacts and 3) has acceptable model metrics. The metrics evaluated are Root Mean Square Error (RMSE), R^2 (Pearson correlation coefficient), and Mean Absolute Error (MAE). Model metrics based on transformed data (log metrics for the log scale linear regression model and the neural network are based on the back-transformed data to ensure a fair comparison between the different models.[169] The physical feasibility of each

component in each model was assessed by applying the model to a theoretical dataset in which each component varies between 1-15 wt%, while the others remain constant.

6.2.4.1.2 Datasets and variables used for modelling

In section 6.2.2 we discussed the selected ingredients blends that are required to cover the full formulation matrix. To answer the main question of this study, which is to assess whether a single model can quantify each techno-functional property of multiple crops, different datasets were modelled:

1. yellow pea data only: ingredient blends made from yellow pea commercial isolates and mildly refined ingredients
2. lupine seed data only: Ingredient blends made from lupine commercial isolates and mildly refined ingredients
3. yellow pea and lupine seed data combined: datasets number 1 and 2 with the addition of ingredient blends from yellow pea and lupine seed ingredients mixed together.

For each techno-functional property, 218 – 508 observations were collected (Table 6.2). Sizes may vary as the datasets were developed individually and some sets were used for protocol development.

Observations were divided into a training and a test set, in such a way that both sets had an equal representation with respect to the composition of protein, starch, fibre (Figure 6.1). This could not be done randomly, since the spread of training and test data had to be even throughout the composition space. In addition, the selected ingredient blends are measured at specific concentrations between 1 and 25 wt%. To check for interpolation, some concentrations within the range are only tested, which is also indicated in Table 6.2. The viscosity dataset of yellow pea ingredients was not obtained in the same systematic way as the other functional properties. Therefore, it does not include specific training and testing concentrations and is rather measured throughout the concentration range.

The main macro components (MC) (protein, starch, fibre, and the rest of the components) represented in concentration (wt%) were always input variables. The output variables were always the techno-functional properties, being the final

viscosity (mPa.s at 160 rpm), the gel stiffness (kPa), foaming capacity (overrun %), and the emulsion separation velocity (µm/s). Potential effects of the processing history on protein functionality were considered by splitting the input variable 'protein' according to its solubility or nativity where possible (e.g., denatured and native protein). This results in three sets of input variables: 1) the MCs, 2) MCs with protein split into soluble and insoluble protein, and 3) MCs with protein split into native and denatured protein. For the models based on all data, the input variables were also split according to crop for some properties (e.g., protein from yellow pea and protein from lupine).

Table 6.2 Sizes of total datasets (n) for each functional property and the percentage of the dataset that was used as a test set. The measured concentrations at which the ingredient blends are trained and tested, as well as only tested are indicated.

	All data	Yellow pea	Lupine	Mix	Concentrations (wt%)	
Functional property	Total dataset n (%test)	Total dataset n (%test)	Total dataset n (%test)	Total dataset n (%test)	Training + test	Extra test
Emulsions	508 (38)	364 (35)	88 (44)	56 (48)	1, 1.5, 2, 2.5, 4, 9, 15, 21	6, 7, 12, 18
Foams	423 (29)	267 (22)	99 (37)	57 (47)	1, 2, 4, 5, 9, 15, 21	7, 12, 18
Gelation	242 (47)	163 (45)	52 (48)	27 (44)	13, 18, 25	15, 21
Heated viscosity	225 (24)	110 (14)	85 (34)	30 (37)	2, 6, 10, 14, 18, 22	12, 16
Unheated viscosity	218 (25)	101 (14)	78 (36)	39 (33)	2-22* 2, 6, 10, 14, 18, 22**	12, 16**

*Measured for yellow pea ingredients blends
**Measured for lupine ingredient blends.

Significant differences in model metric performance between the different sets of independent variables were assessed for models that used a cross validation (regularized polynomial regression, neural network, random forest, and spline regression). The models were run five times and significant differences were assessed using a Two-Way ANOVA and Tukey's post-hoc test ($p \leq 0.05$) to aid in the selection. When a large variation between the repetitions was detected, the model with the smallest test errors and acceptable behaviour was used for variable set selection and further analyses.

6.2.4.2 Dataset size reduction

To reduce the experimental effort, it is of interest to estimate the minimum sample size required to model the selected techno-functional properties and reach the optimal prediction accuracy. The relevance of the size of the datasets is assessed by running models on a subset of observations from the complete dataset. This is only performed on the final model selected based on all data. The number (n) of selected observations ranges from n = 10 (the minimum needed to create a model) to the length of the complete dataset. The selection of observations is done using the Kennard-Stone algorithm, which selects observations with a uniform distribution over the whole compositional space, starting with the indicated n number of observations spread out as possible based on the Euclidian distance.[186]

For all datasets (from n = 10 to the maximum possible number of observations), the prediction accuracy is determined by a (nested) leave-one-out-cross validation (LOOCV).[187] This entails that for each dataset with length n, in an outer loop observation i is taken out of the dataset and the models are trained on n-i observations in an inner loop, using the regular model training and cross validation as described earlier, which is called the inner cross validation. The model cross validation was reduced to two times repeated 5-fold cross validation to reduce computing time. After the models in the inner loop were trained, the left-out observation i in the outer loop was predicted by the model. This is repeated for all n observations in the dataset until all have been left out and predicted. The prediction error (Q^2) is determined by equation 6-3 with y_i being the observed value, $\hat{y}$ the predicted value from the outer loop, and $\bar{y}$ the mean of all observed values.

$$Q^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (6\text{-}3)$$

6.3 Results

6.3.1 Experimental data techno-functional properties

Figure 6.2 shows scatterplots for data obtained for all techno-functional properties selected in this study: emulsion stability, foaming capacity, gel stiffness, and viscosity (heated and unheated) as a function of the composition of the measured ingredient blends.

The gel stiffness ranges from around 0 to 800 kPa and is mainly dominated by the presence of starch, which was also found in another study.[66] The gel stiffness of gels made with ingredients from lupine seeds is lower than with yellow pea and mixtures of both crops due to the low starch concentration in lupine seeds. The heated viscosity ranges from 1 to approximately 45,000 mPa.s and is also mainly dependent on starch gelatinisation.[188] The unheated viscosity of all pastes is more dependent on the capacity of the insoluble material to absorb water, which was found relatively high of fibre and protein in yellow pea in another study for yellow pea, as opposed to starch.[66] The dominating effect of fibre in lupine was also hypothesized to be largely responsible for the viscosity in suspensions.[48] The unheated viscosity is significantly lower than in the heated variant, ranging from 1 to 6,000 mPa.s.

The stability of emulsions is influenced by a combination of interfacial stabilisation as well as viscosity[189]. The emulsion separation velocity of the measured emulsions ranges from 0 to approximately 160 μm/s, with 0 μm/s being completely stable in the theoretical period of one month at 1g. The separation velocity of the measured emulsions is highly dependent on the concentration of the ingredients, mostly the protein in the ingredient blends and regardless of the crop. Subtler differences in the ability of the components to stability interfaces are difficult to observe in these plots and are also not in the scope of this study.

The foaming capacity of the measured foams can be divided into two regimes: 0-2.5 wt% protein and >2.5 wt% protein. In the first regime, the foaming capacity increases with protein concentration. This can be attributed to interfacial stabilisation by proteins and increasing viscosity due to the generally higher concentration of the ingredients.[190] In the second regime, the foaming capacity reaches a plateau or decreases again with increasing concentration of ingredients in the foams. We hypothesized that the latter occurs due to the increase in viscosity, which complicates the incorporation of air in the foams. In some cases, the viscosity becomes so high that the dispersion had no liquid layer (free water) in the measuring tube anymore. As a result, for these solid-like foams there is no or little incorporation of air and therefore the height of the foam cannot be determined anymore. Hence, the foaming capacity is registered as 0%. However, the foam overrun can also be 0% due to a lack

of components that can form a foam in a more liquid foam. To prevent any confusion in the interpretation of a 0% foam overrun in a liquid or solid-like foam, the latter is filtered out of the dataset using a decision tree using their unheated viscosity, which will be explained further in a later section.

6.3.2 Quantification of the techno-functional properties

Due to the large number of results that are generated during the selection of the appropriate model to quantify the relationships, the tables and figures of all models are supplied in the Appendix with a short explanation of why certain models were selected. Table 6.3 summarizes the models that are selected to quantify the techno-functional properties. Figure 6.3 depicts the RMSE of the models based on the three analysed datasets based on 1) yellow pea and lupine seed ingredients separately, 2) both ingredients as well as mixtures of those, and 3) both ingredients and mixtures but with the input variables protein and fibre according to origin. The raw data of the RMSE and also the R^2 and MAE of these models are provided in the Appendix. Figure 6.4 shows the parity for all selected models.

6.3.2.1 Pre-processing foaming dataset

The model for the foaming capacity was generated using only the liquid foams. Liquid and semi-solid foams were classified using a decision tree with the unheated viscosity (quantified in this study) as the predictive variable for the separate crop and combined datasets (Table 6.4). The foams of the individual crops were classified as liquid or semi-solid foams with high test accuracy (>93%) and with a moderate to a strong level of agreement (κ of 0.63 and 0.84 for yellow pea and lupine respectively) with the test set.[185] The κ value indicates the probability of a sample ending up in one category by chance. The combined dataset, which includes yellow pea, lupine and mixtures of those, has a similar test accuracy (94%) and also a strong level of agreement (κ = 0.82). The difference between the datasets can be attributed to the classification threshold, which is dependent on the model of the unheated viscosity and can therefore vary among crops.

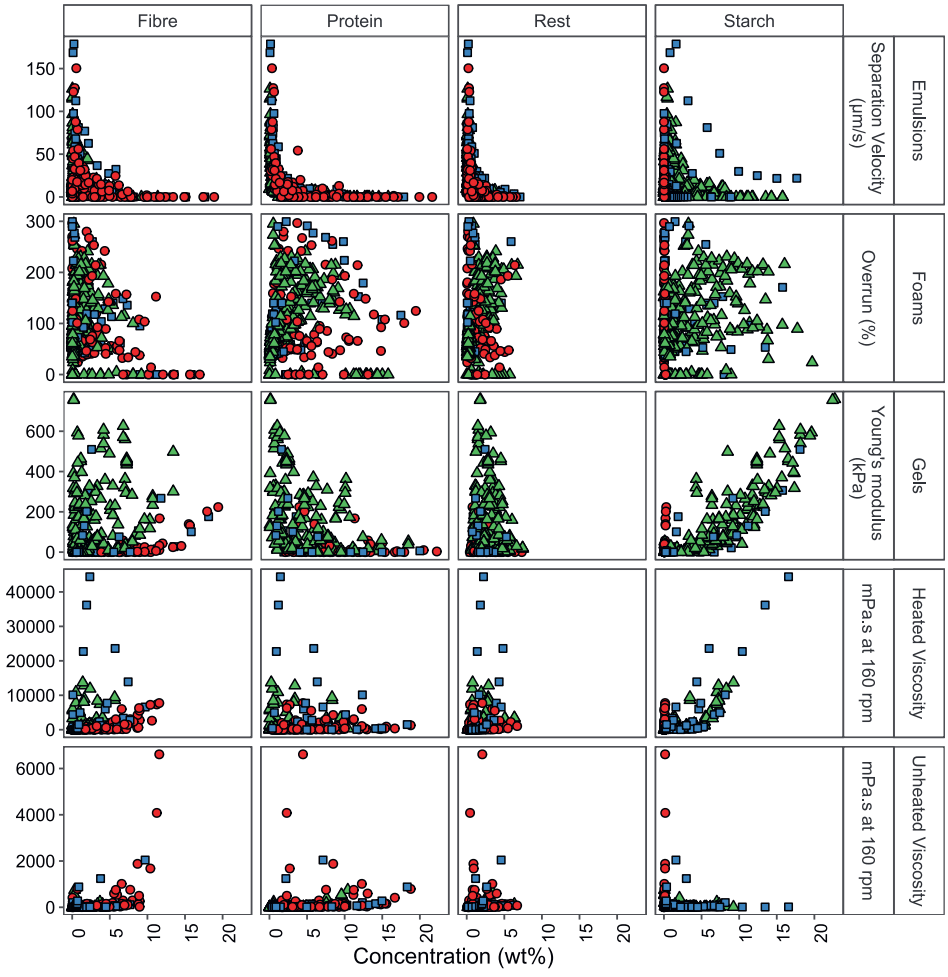

Figure 6.2 Scatter plots showing the relation between the techno-functional properties and the composition (protein, starch, fibre, and rest concentration) of the ingredient blends (the colour and shape indicate the crop: yellow pea▲, lupine ●, and mixtures of both ■).

Table 6.3. Summary of models selected to quantify the relation between the ingredients and techno-functional properties with the main macro components (MC) or with the main macro components with the protein split according to solubility (SL) as input variables. The split column indicates whether models containing data from all crops included a split in independent variables according to crop.

Functional property	Crop	Best model	Variable set	Split
Emulsion stability	Yellow pea	Random forest	SL	-
	Lupine	Neural network	MC	-
	All	Random forest	MC	Yes
Foaming capacity	Yellow pea	Random forest	MC	-
	Lupine	Random Forest	MC	-
	All	Random Forest	MC	Yes
Gelation	Yellow pea	Neural network	MC	-
	Lupine	Neural network	MC	-
	All	Neural network	MC	No
Heated viscosity	Yellow pea	Spline regression	MC	-
	Lupine	Neural network	SL	-
	All	Spline regression	MC	No
Unheated viscosity	Yellow pea	Regularised polynomial linear regression	MC	-
	Lupine	Neural network	MC	-
	All	Neural network	MC	Yes

Table 6.4 Performance of classification of liquid and semi-solid foams

Crop	κ - Train	Accuracy Train (%)	κ - Test	Accuracy Test (%)	Classification threshold (mPa.s)
Yellow pea	0.85	97	0.63	93	226
Lupine	0.94	98	0.84	95	412
All data	0.82	96	0.82	94	158

6.3.2.2 Models fitted on datasets from individual crops

The relations for each techno-functional property based on datasets from individual crops were quantified using a regularized polynomial linear regression, spline regression, random forest, and neural network. The models fitted from yellow pea ingredients for the emulsion stability as well as the lupine ingredients for heated viscosity benefited from splitting the independent variable protein according to their solubility. A considerable number of models have a stable prediction, which means that the train and test errors are close to each other, for example with the heated viscosity and emulsion stability. In some cases, the test error is considerably higher than the training error, for example in the case of foaming capacity. The parity plots in Figure 6.4 indicate that the foaming capacity is overestimated at lower concentrations and underestimated at higher concentrations for yellow pea and (especially) lupine ingredients. The plotted behaviour of the random forest for foaming capacity also show discontinuous behaviour, which is undesired. Yet, there is at this point no better alternative (Appendix). Besides the foaming capacity, the parity plots depict that all other techno-functional properties are still quantified within an acceptable range using the selected models for the individual crops.

6.3.2.3 Models fitted on combined dataset

Next, the observations of yellow pea and lupine, as well as mixtures of those were combined and modelled together. As the same component from different crops can have different contributions to the final functional properties, we considered whether we would need to split protein and fibre according to their origin as input variables (e.g., protein from yellow pea and lupine).

The error of the prediction from some models fitted on all data combined is in the same range as the error from the models fitted on data from single crops only (Figure 6.3). For example, in the case of emulsion stability of yellow pea ingredients or foaming capacity of both datasets of yellow pea and lupine ingredients. For some models, the errors are considerably worse when fitting models based on combined datasets from yellow pea and lupine ingredients. Splitting protein and fibre according to origin improved the prediction error in some cases, for example for the emulsion stability of lupine ingredients as well as the unheated viscosity of yellow

pea ingredients. The errors of the models based on the combined datasets are now in almost all cases in the same range as the models based on the individual datasets, with or without split of protein and fibre according to their origin. This is not the case for the heated viscosity, for which the error is substantially higher when modelling a combined dataset compared to an individual. This is attributed to the poor prediction of observations that were based on mixtures of yellow pea and lupine ingredients.

Based on the model metrics, it is a logical next step to model the combined dataset with a split according to the components' origin for the emulsion stability. For the remaining techno-functional properties, this is not straightforward since the model metrics do not consistently improve when predicting all three datasets using combined models with a split. To facilitate the decision on a final model, the behaviour of each component in each model is evaluated. The trends of the fitted individual models were revealed using a theoretical dataset, in which one component increases, while the other components remain constant (Figure 6.5). We focus only on protein and fibre since starch is only present in considerable concentrations in yellow pea ingredients. From this Figure, it can be derived that protein and fibre in lupine have a very different contribution to the foaming capacity, and to a lesser extent for unheated viscosity and emulsions stability. Protein and fibre from yellow pea and lupine have a similar contribution to the gel stiffness and heated viscosity. Therefore, the selected combined models for quantifying the emulsion stability, foaming capacity, and unheated viscosity will be based on independent variables that are split according to crop origin. The behaviour of the models based on all data (split or without split) is also plotted in Figure 6.5. The parity plots of the final models fitted to the combined dataset, also show that the residuals lay in the same range as from the models based on individual datasets.

In summary, models can be fitted such that they predict the selected techno-functional properties of both yellow pea and lupine samples and mixtures of those although with sometimes a reduction in model accuracy. The decision on the best final model is also not always straightforward as there can be multiple models or sets of variables that are in the same range regarding the metrics. Therefore, selection of

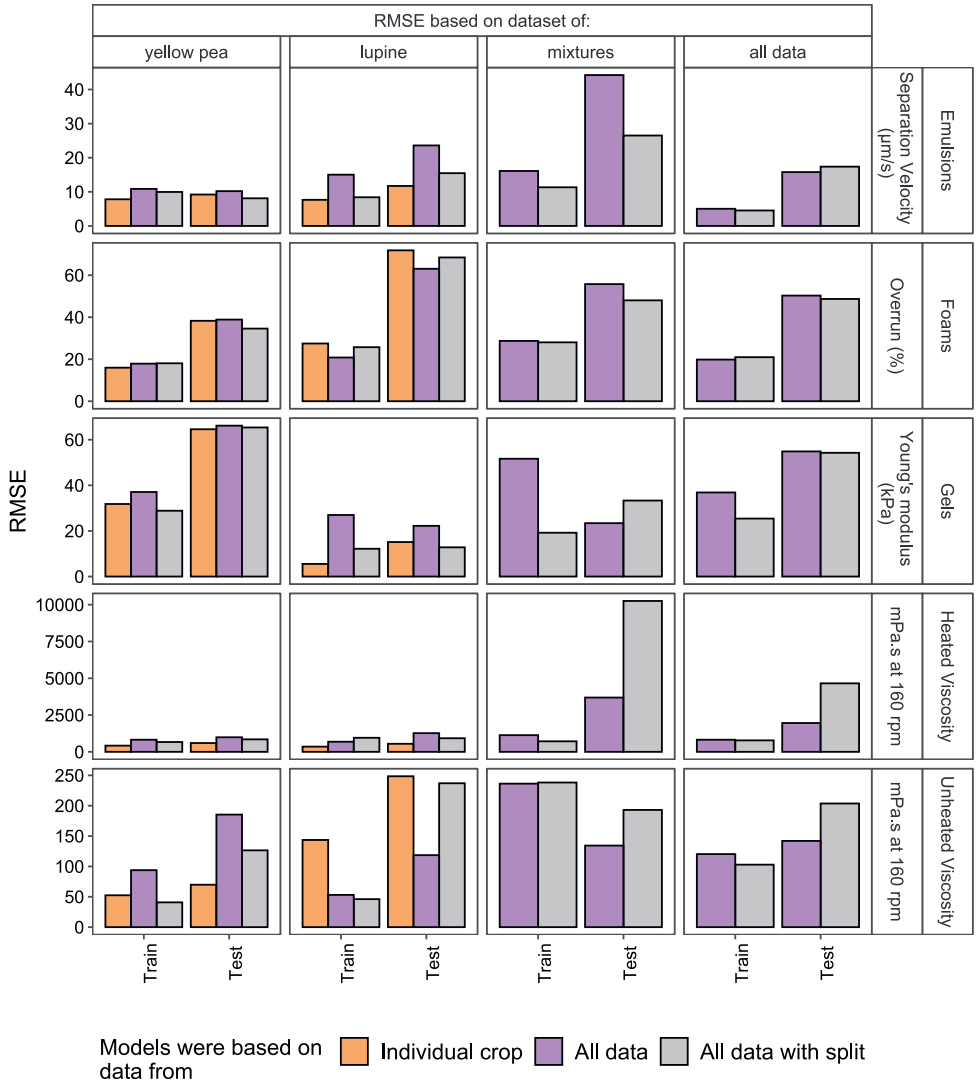

Figure 6.3 RMSE of the models based on data from individual crops (yellow pea and lupine) and on all data (yellow pea, lupine, and mixtures of those) and on all data with a split in protein and fibre according to origin, calculated for the samples of yellow pea, lupine, mixtures, and all data together.

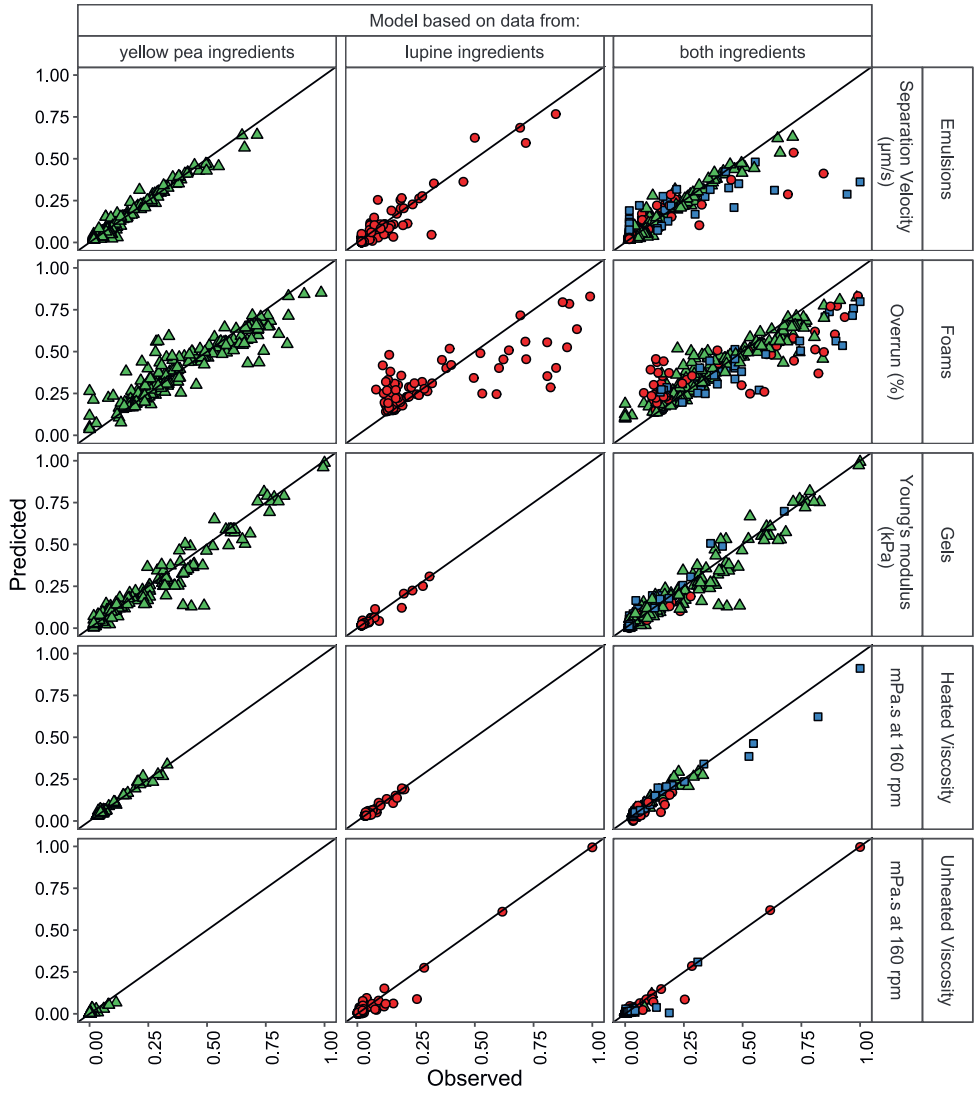

Figure 6.4. Normalized parity plots for three models based on yellow pea, lupine, or all ingredients, visualized for each functional property. the colour and shape indicates the origin of the sample: yellow pea▲, lupine ●, and mixtures of both ■.

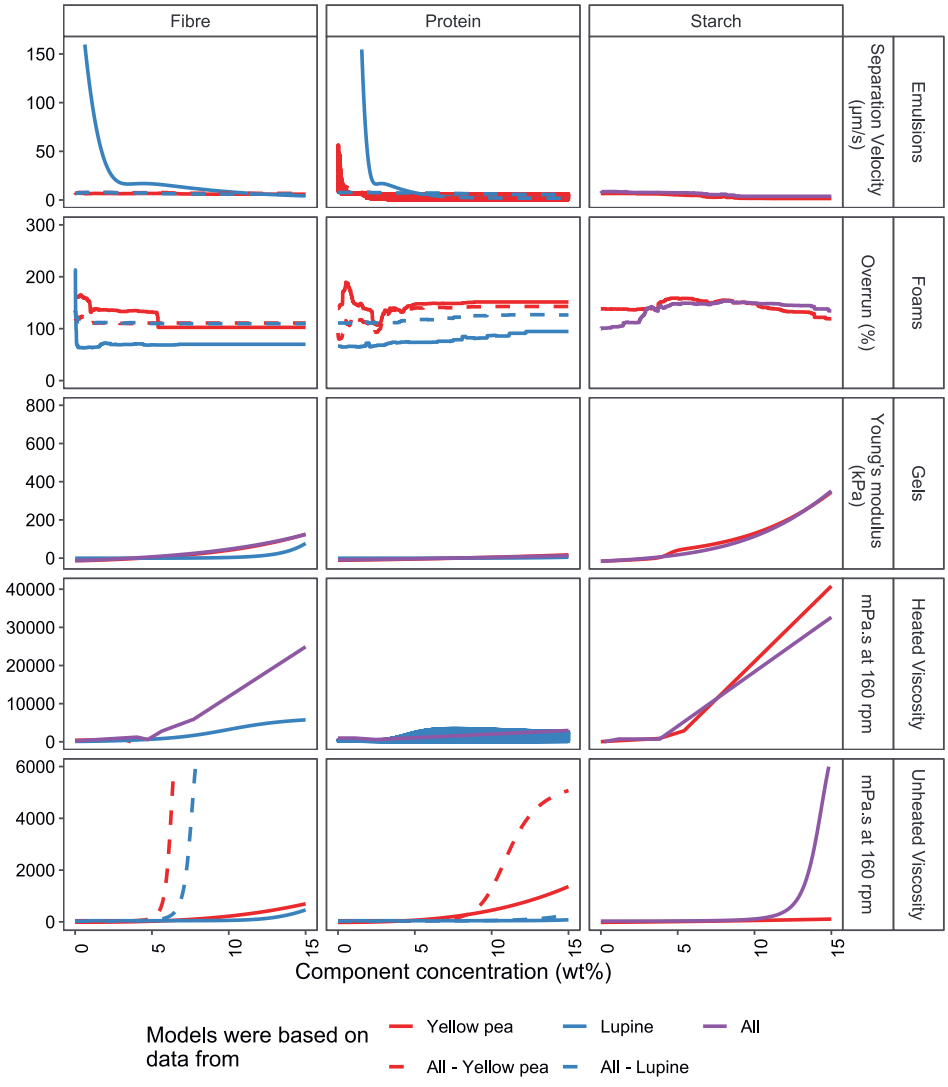

Figure 6.5 Behaviour of each component in the fitted models based on yellow pea, lupine, or all (with or without split in protein and fibre). The models were used to predict theoretical datasets in which each component increases from 1-15 wt% while the remaining component are set constant to 2 wt%. In case a model required the soluble and insoluble protein is as input variables, the solution space is given as shaded area.

.

the final model is a trade-off between having low model errors as well as feasible behaviour of the components. The selected models are also highly dependent on the training dataset and can therefore change with new data.

6.4 Discussion

6.4.1 Reflection on the predictive variables

Most of the selected techno-functional properties could be described using the main macro components (protein, starch, fibre, and the rest) as input variables, for both the models based on yellow pea and lupine ingredients individually plus the combined datasets. This implies that the effect of the processing history as such on the state of the protein and functional property is negligible compared to other components or effects (i.e., interactions). For gelation and heated viscosity this was expected due to the dominance of starch gelatinisation.[188] We did expect differences between the effect of the degree of processing (i.e., conventionally or mildly fractionated) on protein in the foaming capacity, emulsion stability, and unheated viscosity. We hypothesize that due to the large range of measured components and hence, the large range of for example foaming capacities, these differences become negligible. Another effect may be that to measure some properties, the product is heated. This eliminates any differences in the degree of denaturation of the proteins, in case of yellow peas above about 83°C.[66,99] In addition, the protein from the mildly refined protein-rich fraction is only present at lower purity (<55 wt%), and therefore at these low concentrations other components and/or their interactions could be dominant.

The models that were fitted on the combined datasets of yellow pea and lupine ingredients as well as mixtures of those had in most cases a similar prediction error to the models based on individual crops. Based on the model metrics and the behaviour plots, we concluded that the foaming capacity, emulsion stability, and unheated viscosity required a separation of protein and fibre according to the source. Since these three properties were all based on unheated powders, it is no surprise that these three require the same split. In contrast to the heated viscosity, the unheated viscosity is mainly dependent on the size of the protein aggregates[11] and the water holding capacity of the protein[191]. With an increase in fibre the amount of

free water is also reduced, which increases the viscosity.[66] Since these properties can differ among different legumes, a split may be required. Both heated viscosity and gelation result from heated dispersions, starch gelatinisation is dominant. As starch is present only in very low concentrations in lupine, it is hypothesized that there is no need for a split in components according to the source.

6.4.2 Quality of the model predictions

Although the training errors from the fitted models are within a reasonable range, the test errors are considerably higher for some functional properties. While it is common that the test error is larger than the training error[28], a large difference could point to an overfitted model. For example, the RMSE of the test set of emulsion stability is more than twice as large as the RMSE of the training set. This means that the trained models may not be suitable to generalize to other unseen data. This can be the result of having too many predictive independent variables or having too complex models (e.g., too many neural nodes or high degree polynomial)[28,173]. Another reason could be that there is not enough data to train the model properly.[192] Having too complex models is most likely not relevant since we carefully reduced the model complexity by using cross validation during model training. The latter reason could be more relevant in this case. Although the total sizes of the datasets are fairly large, the test sets contain in some cases up to 40% of all observations. This leaves us with a relatively sparse training. Therefore, we evaluated whether the size of the dataset is sufficient.

We assessed the model performance of data subsets with sizes ranging from 10 to the total number of observations that were measured in this study. Due to the small size of the datasets, we applied a nested leave-one-out cross-validation (LOOCV). In this way, every sample becomes a test sample one time while the other observations are used to train a model. The model performance was expressed with Q^2, which is similar to the R^2 of the test value but calculated with the equation that expresses the explained variance of the residuals (Equation 6-3). The LOOCV can be known for overfitting due to the large training set as opposed to the test set.[193] As a control, we have plotted the Q^2 of the original division in the training and test set (up to 40% test set) from this study.

Figure 6.6 shows that with the addition of more observations the prediction quality of emulsion stability and foams keeps on increasing. This means that indeed the dataset size for these was not sufficient. At around 227 observations, there is a sudden decline in Q^2. A closer look into the additional observation added here (observation nr. 227) shows that this point is related to a relatively high emulsion separation velocity of 168 µm/s, while the observations before this point had a maximum separation velocity of 112 µm/s. We expect that this magnitude of separation velocity is undertrained and therefore reduces the overall accuracy of the model. After the addition of the 227th point, the accuracy starts to improve again. Therefore, besides having a homogenous distribution of the compositional space, the distribution of the output variables should also be homogenous.

The Q^2 of the prediction for gel stiffness as well as heated and unheated viscosity reaches a plateau after around 100 observations. As the Q^2 for the prediction of gel stiffness and heated viscosity is quite high already (> 0.8), we conclude that we do

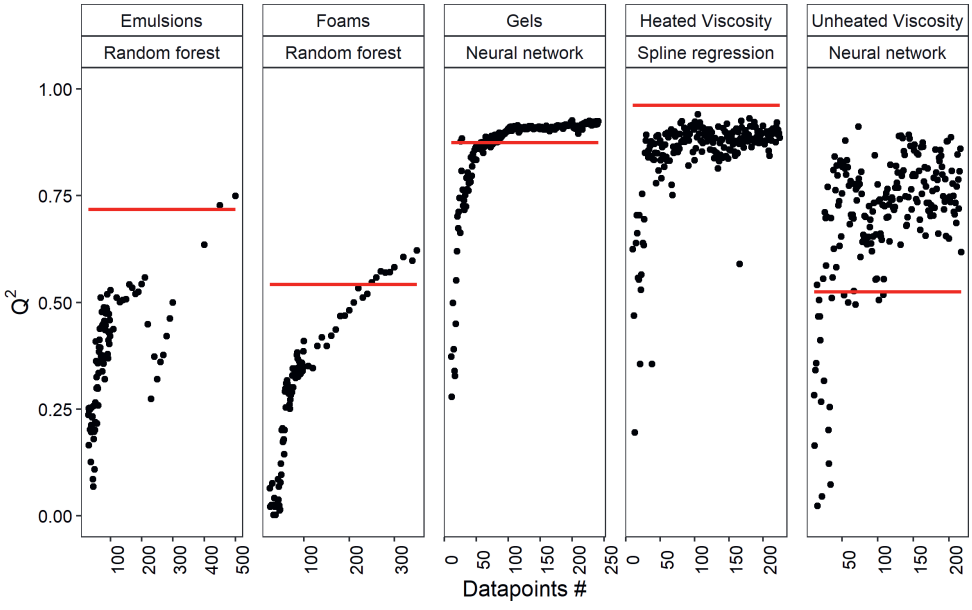

Figure 6.6 Accuracy (Q^2) of models generated for each functional property with a dataset size ranging from 10 to the maximum number of observations measured. The red line indicates the Q^2 of the original test dataset used in this study.

not always need the complete dataset to achieve this model performance. If the train error is still a lot lower than the test error, this could be the result of unsuitable or too many predictive variables. The Q^2 of the unheated viscosity reaches a plateau, yet the variation of the Q^2 is large. This could indicate that the data of the viscosity is not spread homogenously either among the compositional space or in terms of the output, in this case, final viscosity. Since the plateau from the LOOCV is slightly higher than the one of the original dataset, there is a chance that the LOOCV slightly overestimates the Q^2. This means that the Q^2 of the plateau in fact could be slightly lower.

In summary, some overfitting takes place in the fitted models in this study which can be attributed to a lack of samples in the case of emulsion stability and foaming capacity. The analysis for emulsion stability also showed that besides having well distributed data over the composition space, it is also important to have a proper distribution in the model output, in this case, separation velocity. For the other properties overfitting can be attributed to for example the selection of predictive variables. We suggest that when measuring the functional properties of a new crop, analyses like these should be executed in parallel to the experimental measurements, such that the experimental work can be stopped as soon as the plateau is reached. This could ultimately be paired with an automated sampling method.

6.5 Conclusions

Techno-functional properties that are important in food product formulations were quantitatively modelled using machine learning algorithms. These properties were the thickening behaviour (viscosity and gelling), emulsions stability, and foaming capacity. We evaluated whether a single model could predict these properties from multiple crops. The models were therefore fitted on individual datasets from yellow pea and lupine ingredients as well as on a combined dataset that contained both ingredients from yellow pea and lupine as well as mixtures of those two.

The relations between ingredients from single crops and their functional properties were described using a regularized polynomial linear regression, spline regression, random forest and neural network. The latter three models were selected to predict

the functional properties from all data of yellow pea, lupine, and mixtures of those. For some properties, it was necessary to split the protein and fibre according to their sources to obtain better model performance and physically plausible behaviour of each component. The test errors of the models based on the combined datasets were in the same range as those of the models based on individual datasets. Only the heated viscosity was predicted worse using a model based on combined data. Some overfitting was detected in the fitted models, which was later attributed to a lack of samples or the selected predictive variables. This analysis indicated as well that not all data were necessary predict all properties.

This study demonstrates the possibility to extend the functionality-driven selection of food products to multiple crops and functional properties and presents a method to evaluate the relevance of increasing the size of the dataset. The presented method aids in designing food formulations that may help in significantly reducing the overall chain environmental impact of the production of the foods, by reducing the amount of refining necessary to acquire the right product properties.

6.6 Acknowledgements

Anja Schröder, Vincent Rang, and Xilong Zhou are acknowledged for their help with the method development for emulsions and Jack Yang and Bente Sommers for foams.

6.7 Appendix

6.7.1 Ternary plots with rest fraction

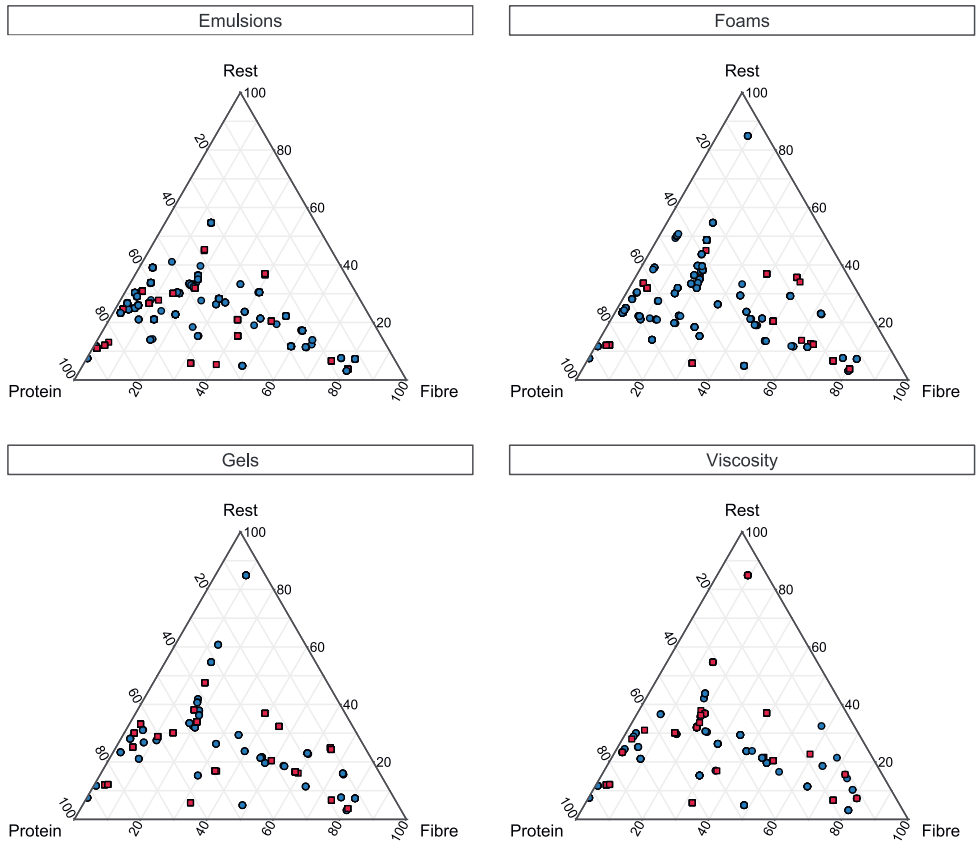

Figure A 6.1 Ternary plots for samples (ratio protein, starch, and rest) of all techno-functional properties with train (●) and test (■) observations All determined in a range of 1-25 wt%.

6.7.2 Example of model selection unheated viscosity of lupine ingredients

From the model metrics it is derived that the neural network is the best candidate for quantifying unheated viscosity of lupine ingredients. The behaviour scatterplots shows physically plausible behaviour in terms of trends and negative values for the neural network. There are no significant differences between the test metrics between the base components and the split according to solubility. Therefore, the best model is the neural network with main macro components. As there is quite a large variation in the five times repeated neural networks, the one with the lowest test error is chosen for further analyses.

Table A 6.1 Model metrics (MAE, RMSE, R^2) models for quantifying unheated viscosity with main macro components as independent variables for yellow pea ingredients.

Model	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Linear Model	494.73	0.80	297.56	482.37	0.59	333.00
Log Linear Model	721.84	0.59	179.01	295.52	0.65	94.52
Poly Model	212.58	0.96	158.54	349.74	0.78	215.63
Regularisation Model	445.65	0.84	214.43	467.02	0.64	278.70
Spline Model	220.66	0.96	124.12	352.40	0.68	201.38
Random Forest	308.43	0.95	92.37	344.78	0.80	130.84
Neural network	143.59	0.98	77.99	248.52	0.92	115.63

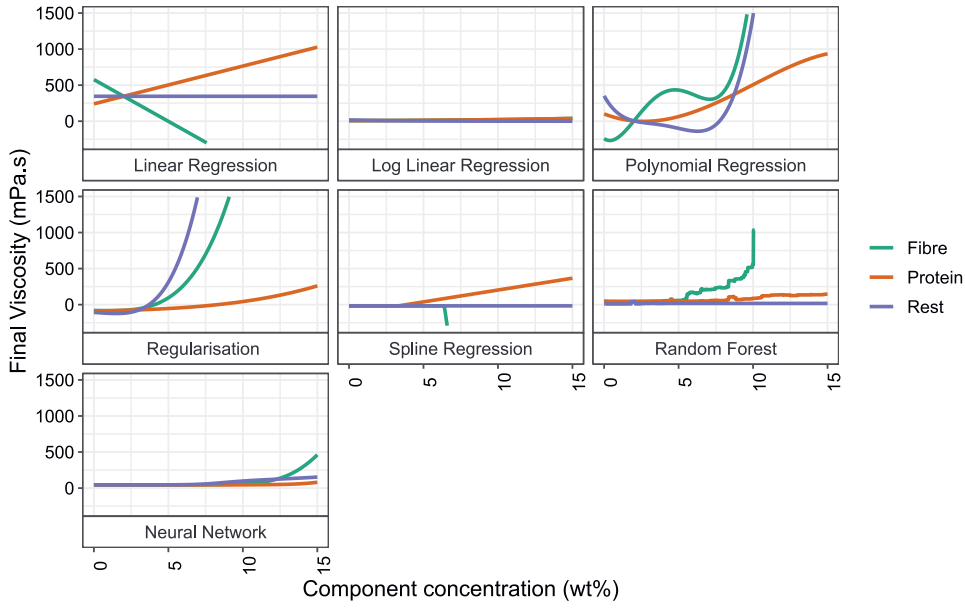

Figure A 6.2 Scatterplot of the behaviour of each component in the evaluated models for quantifying the unheated viscosity of lupine seed ingredients with the main macro components as independent variables. The composition of each component increases from 1-15 wt% while the other stay constant at 2%.

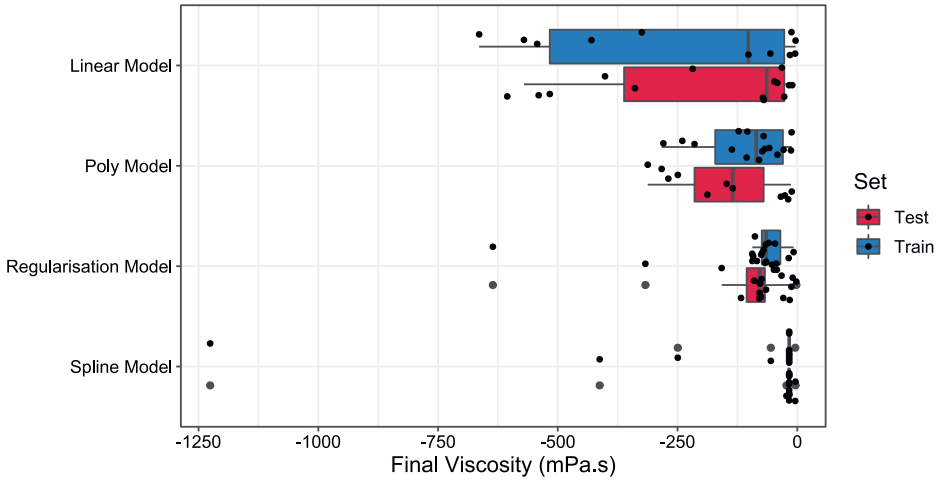

Figure A 6.3 Boxplot of negative values predicted by the evaluated models to for quantifying the unheated viscosity of lupine seed ingredients with the main macro components as independent variables.

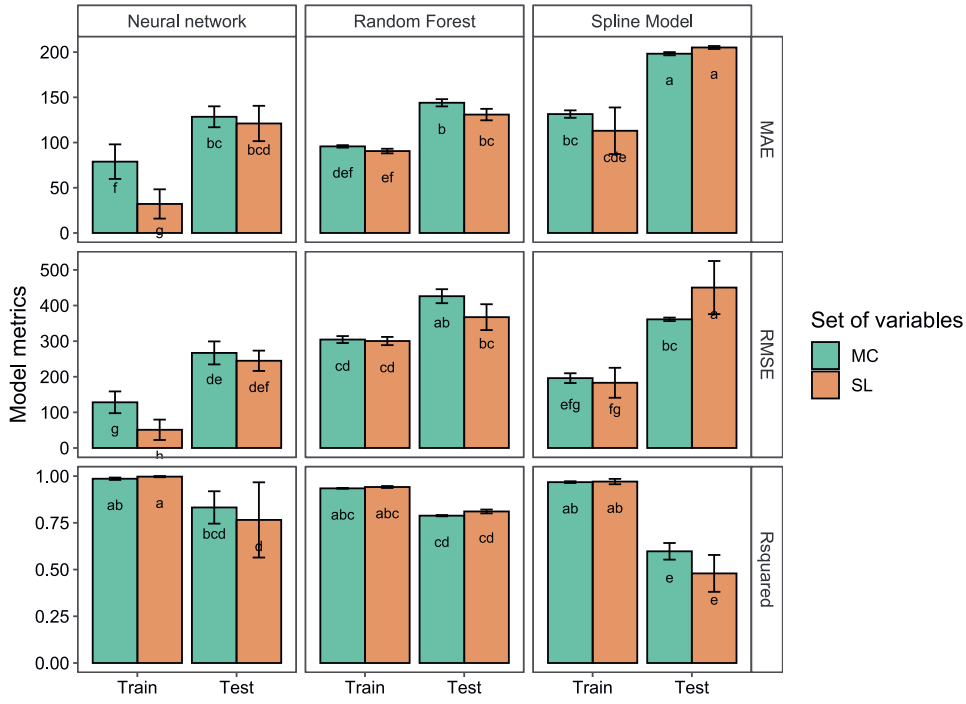

Figure A 6.4 Bar chart containing the model metrics (MAE, RMSE, R^2) to predict the unheated viscosity (mean absolute error (MAE), root mean square error (RMSE), and R^2) generated five times for the neural network, random forest, and spline regression for yellow pea ingredients with the main macro components (MC) and main macro components with a split according to soluble protein (SL) as independent variables. Letters indicate a significant different (P<0.05).

The supplementary information is stored temporarily at the following link *https://figshare.com/s/b26eceed43dec72cd717* or use QR until published.

6.7.3 Model metrics and parity plots of models based on all data, visualised per crop.

Table A 6.2 Summary with model metrics (MAE, RMSE, R^2) of models selected to quantify the relation between the ingredients and techno-functional properties with the main macro components (MC) or with the main macro components with the protein split according to solubility (SL) as input variables. The split column indicates whether models containing data from all crops included a split in independent variables according to crop.

Functional property	Crop	Best model	Variable set	Split	RMSE Train	R^2 Train	MAE Train	RMSE Test	R^2 Test	MAE Test
Emulsion stability	Yellow pea	Random forest	SL	NA	3.03	0.99	1.6	4.3	0.88	2.71
	Lupine	Neural network	MC	NA	7.66	0.9	5.21	11.72	0.87	7.25
	All	Random forest	MC	Yes	4.53	0.97	2.49	17.38	0.62	7.72
Foaming capacity	Yellow pea	Random forest	MC	NA	15.96	0.94	11.7	38.29	0.65	26.84
	Lupine	Random Forest	MC	NA	27.48	0.9	17.95	71.83	0.18	56.77
	All	Random Forest	MC	Yes	20.16	0.94	14.02	47.66	0.55	35.2
Gelation	Yellow pea	Neural network	MC	NA	31.82	0.97	22.57	64.6	0.88	42.05
	Lupine	Neural network	MC	NA	5.54	0.99	3.27	15.14	0.92	7.52
	All	Neural network	MC	No	36.88	0.96	26.7	54.87	0.89	33.63
Heated viscosity	Yellow pea	Spline regression	MC	NA	420.6	0.98	234.3	597.53	0.95	457.04
	Lupine	Neural network	SL	NA	353.93	0.96	224.5	547.07	0.93	280.73
	All	Spline regression	MC	No	822.83	0.94	554.89	1960.03	0.96	1076.79
Unheated viscosity	Yellow pea	Regularized polynomial regression	MC	NA	52.47	0.71	21.95	69.86	0.99	36.05
	Lupine	Neural network	MC	NA	143.59	0.98	77.99	248.52	0.92	115.63
	All	Neural network	MC	Yes	102.93	0.98	32.29	203.66	0.79	110.86

6.7.4 MAE and R^2 of all data versus individual

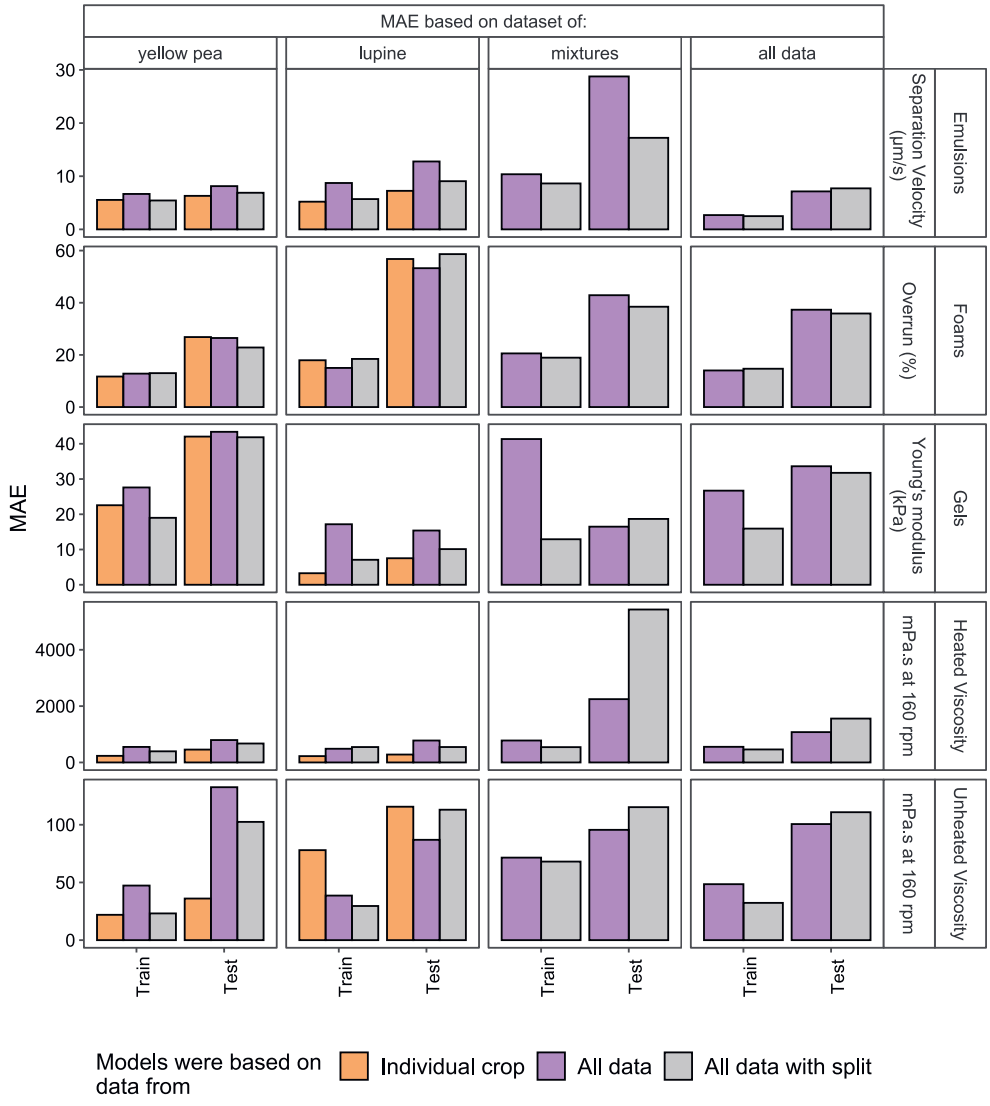

Figure A 6.5 MAE of the models based on data from individual crops (yellow pea and lupine) and on all data (yellow pea, lupine, and mixtures of those), calculated for the samples of yellow pea, lupine, mixtures, and all data together.

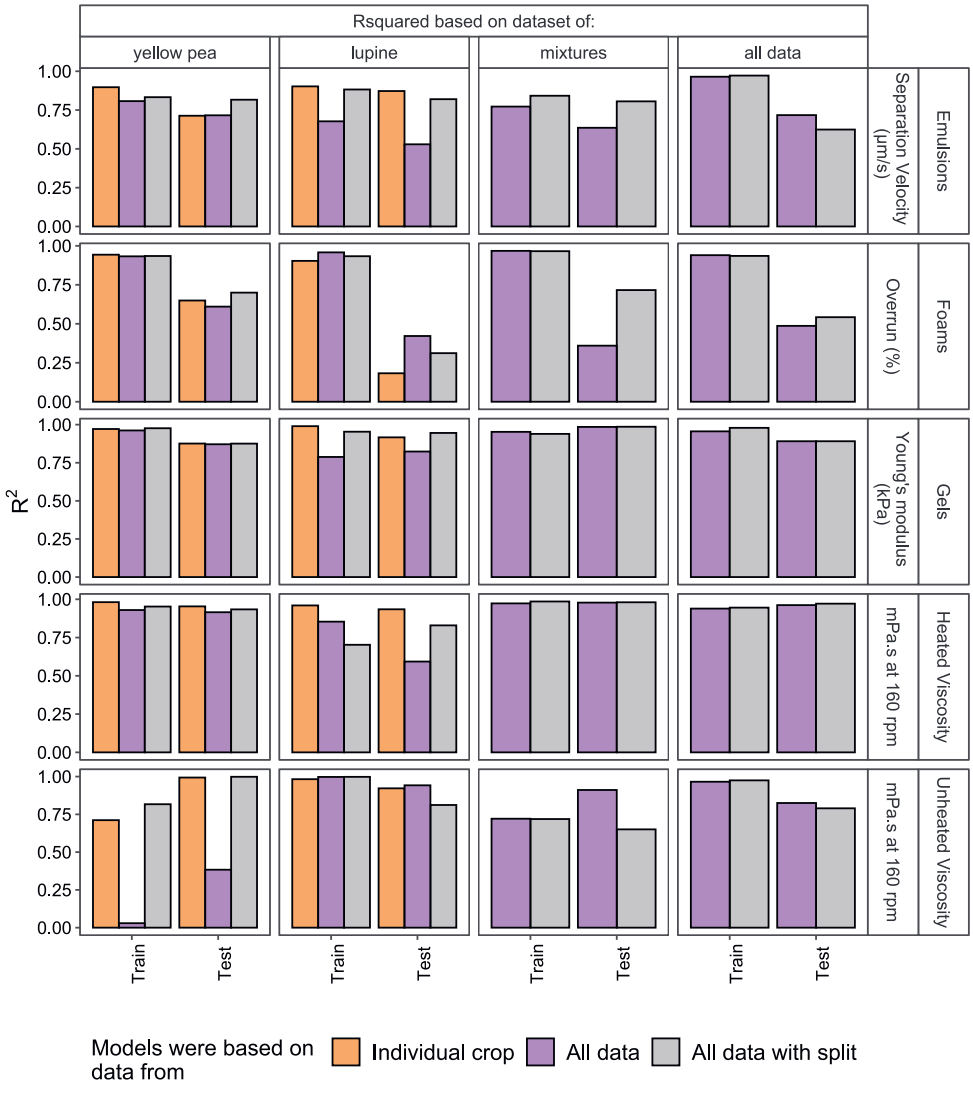

Figure A 6.6 R^2 of the models based on data from individual crops (yellow pea and lupine) and on all data (yellow pea, lupine, and mixtures of those), calculated for the samples of yellow pea, lupine, mixtures, and all data together.

Chapter 7

General discussion

Ingredient formulations driven by techno-functional properties

---

A. Lie-Piang, R.M. Boom, A. van der Padt

Partially submitted

Food ingredients that texturize products like vegan or ready-to-eat foods are conventionally produced by fractionating a crop (e.g., wheat or peas) into pure isolates such as a protein or starch isolate. In the case of protein, flour is often dispersed and subsequently subjected to alkaline extraction, iso-electric precipitation, and drying with for example spray drying. This process route is resource use intensive. In addition, the separation of crops into pure components and later assembling these components again in food products is not efficient, as it recovers only part of the nutrients in the raw material.

The development of fractionation techniques for food ingredients that omit the use of water and/or chemicals has resulted in the development of so-called milder refined ingredients. These processes require much less energy and water and can be more efficient in raw material usage. The ingredients are produced from the same crops but are not as refined, or pure, as conventionally produced ingredients. The impurity complicates the formulation processes of combining several ingredients into a consumer product. Nevertheless, less refined ingredients have high functionality: they can just as well give high viscosity, gel strength or other properties as refined ingredients can.

Previous work indicated that the use of mildly refined ingredients is facilitated by matching the compositions of foods to blends of these ingredients.[18] This resulted in a significant reduction in water and energy usage in the production chain but would not reduce the number of raw materials. The underlying hypothesis of this thesis was that by matching products to ingredients based on the techno-functionalities instead of composition, one can reduce the use of raw materials as well as the use of energy and water. The aim of this thesis was therefore to assess to what extent the use of a functionality-driven selection of ingredients may lead to more sustainable food production.

A case study was used in which yellow pea and lupine seeds are both conventionally and mildly fractionated, after which the ingredients were matched to a product portfolio. As there are many functional properties of ingredients, four main texturizing properties were selected in this study to demonstrate the principle. Models were constructed that relate the composition of the ingredients and the

(techno)functional properties that they provide in consumer foods. These models were then used to minimise the environmental impact of food production while retaining the required textural properties.

This chapter summarizes the main findings of this thesis. The findings are subsequently compiled into an assessment of the effects of the functionality-driven ingredient selection on the sustainability of a final portfolio of 14 products. Lastly, we discuss areas for improvement of the proposed approach.

7.1 Milder fractionation processes and their environmental impact

An overview of milder fractionation methods was provided in **chapter 2 – state of the art.** Although a widely accepted and clear definition of mildness in fractionation methods is not available, in this chapter we discussed all methods that use milder refining steps. These are milder as opposed to conventional processes to extract carbohydrates, oils, and proteins from crops. These are some of the most important macronutrients and their production is resource use intensive. The conventional way to extract oils is by disrupting the seed matrix through milling, mechanical expression, and solvent extraction to isolate the oil. Both elevated temperatures and the use of solvents like hexane degrade the oil and the remaining matrix, for example by denaturation, oxidation. and crosslinking. In a milder alternative, the oil-bearing intracellular vesicles, oleosomes, are extracted whole. Therefore, they maintain their functional properties, protecting the oil against degradation. These oleosomes form natural emulsions and can be directly used in products such as creams. In the conventional method to extract protein, the starch and fibres are first removed with (multiple) centrifugation or hydrocylone steps. The proteins are subsequently solubilized at high pH, isolated by iso-electric precipitation, and then dried into a powder. The other fractions are concentrated and dried as well.

Subsequently in **chapter 2 – state of the art**, both wet and dry milder alternatives to the conventional methods are discussed. Wet alternatives avoid drying and apply for example ultrafiltration. The alkaline extraction and/or iso-electric precipitation step can also be omitted to reduce resource use and maintain the native functionalities of the protein. Dry fractionation completely omits the use of chemicals by separating flour based on particle size and density with air classification

or with electrostatic separation based on triboelectric charging properties. The benefit of dry fractionation is that it is purely a mechanical separation and therefore the product does not need to be dried. Therefore, fractions maintain their native functionality, although at the cost of lower purity.

To assess the difference in sustainability of ingredient portfolios based on extensively or mildly refined ingredients, the environmental impact of the ingredient production from yellow pea and lupine was quantified in **chapter 3 – sustainability**. Ingredients with a lower degree of refining indeed have a lower environmental impact, even when expressed per kg of protein. The latter is important since for example air classified yellow pea concentrate has a lower protein concentration than the conventional isolate. The impact of processing was in some cases similar to the impact of the cultivation of the crops. This makes it important to consider both while assessing the sustainability of these food ingredients. The impact of the cultivation is also dependent on the type of processing, as more efficient routes need fewer materials to obtain a similar amount of ingredient.

7.2 Quantification of techno-functional properties

A compositionally-driven ingredient selection was compared to a functionality-driven selection. For this, the case study was narrowed down to the use of conventionally isolated and air classified ingredients from yellow peas and lupine seeds. A preliminary study was done to conceptualize the approach using the thickening capacity of yellow pea ingredients in **chapter 4 – concept.** The concentration of the main macro components (i.e., protein, starch, and fibre) from yellow pea ingredients could be used to quantify and predict the final viscosity using multiple linear regression. The effect of the processing history on the functionality of protein was found to be negligible to predict the final viscosity. This was attributed to the low protein content of the protein-rich concentrate obtained through air classification as compared to the isolate and the dominating effect of starch in the development of viscosity. With the use of a linear regression model, a window of formulations that resulted in the same viscosity could be identified. Subsequent optimization then led to an ingredient formulation within this window that has a minimal environmental impact. With this approach, food assemblers can select

ingredient formulations with preferred composition and functionality while minimizing the environmental impact.

This approach was extrapolated to other techno-functional properties that are required to texturize the products selected in the product portfolio: gelling, foaming capacity, and emulsion stability. Multiple linear regression was not adequate to model these non-linear functional properties and resulted in poor and physically unfeasible predictions. Therefore, in **chapter 5 – framework**, a framework is proposed to select an appropriate machine learning method for quantifying these types of relationships, while applying expert knowledge of known physical phenomena regarding these properties. The conditions for an appropriate method were physically feasible predictions and acceptable model metrics, such as the correlation coefficient (R^2) and the test errors. Since one machine learning method could not always comply with all conditions, linear regression, random forest, spline regression, and neural network were explored. The framework was illustrated using the gel stiffness of yellow pea ingredients, which was best described by a neural network. A main finding was that there is no general method that can quantify all relationships since machine learning models can easily generate physically unfeasible solutions or show artefacts. Instead, the systematic framework as proposed in this chapter may lead to the optimal model choice for the respective dataset.

So far, we have focused on only ingredients from yellow peas. Nonetheless, ingredient formulations normally contain ingredients from multiple crops. The case study presented in this thesis also includes ingredients from lupine seeds. It was desired to create models that can predict the techno-functionalities of both yellow pea and lupine seed ingredients. Therefore, in **chapter 6 – mixed crops** we assessed the potential of quantifying each selected techno-functional property for both crops using a single model. The heated viscosity and gelation were modelled based on a merged composition of lupine and yellow ingredients. This means that the protein, starch, and fibre from both crops could be combined into single predictive variables. This resulted in the case of heated viscosity in slightly higher test errors compared to the models based on individual crops. For foams, emulsions,

and unheated viscosity, the behaviour of protein and fibre from both crops was significantly different. Therefore, the input variables had to be split according to their source (e.g., protein and fibre from yellow pea and lupine) to achieve acceptable predictions and model behaviour. It was also found that the prediction accuracy of foams and emulsions would benefit from more observations. However, for the remaining properties, the complete measured dataset was not necessary: there is an upper limit to the accuracy of the model at sometimes half of the measured size of the dataset. This also shows that with the current predictive variables, it is not possible to achieve better accuracy with the addition of more observations. Therefore, we are at the upper limit of what can be achieved. Additional or different independent variables would have to be added to further improve the accuracy.

Overall, the functional properties of yellow pea and lupine ingredients could be well modelled for both extensively and mildly refined ingredients using machine learning algorithms. These models predict physically feasible solutions based on the datasets in this study.

7.3 Functionality-oriented selection of ingredients

The models created to describe the functional properties were subsequently used to formulate ingredients for a product portfolio (Figure 7.1). For the current study, this means that both extensively and mildly refined yellow pea and lupine ingredients were considered and could be formulated into a product portfolio. More specifically, a yellow pea protein (YPI), fibre (YFI), and starch (YSI) isolate, as well as a lupine protein (LPI) and fibre (LFI) isolate were used. The milder refined ingredients were a dry fractionated (using air classification) yellow pea flour (YPF), protein-rich fine (YFF) and starch-rich coarse (YCF) fraction, lupine flour (LF), and protein-rich fine (LFF) and a fibre-rich coarse (LCF) fraction. The formulations were minimized based on global warming potential (GWP), water usage, and raw materials.

The selection of the ingredients within this space was based on three scenarios:

1. Base scenario: only isolates are allowed to be used in the portfolio and are matched on composition

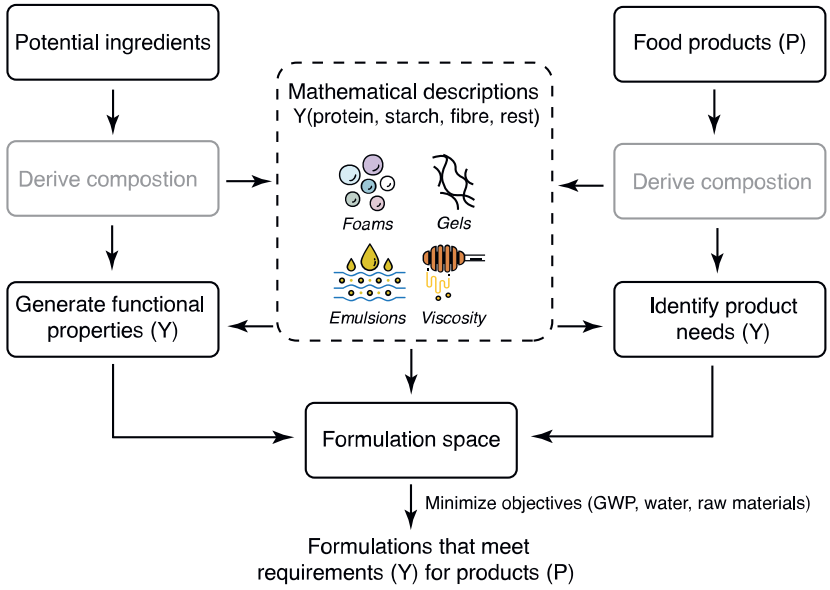

Figure 7.1 Schematic representation of the case description. The composition from the potential ingredients and the food products is first translated into their techno-functional properties using the mathematical descriptions from Chapter 6. The functional properties from the ingredients and products are matched into a formulation space. These are minimised for their environmental impact.

2. Compositional Match scenario (CM): isolates and mildly refined ingredients are allowed in ingredients formulations and are selected based on composition
3. Techno-functional Match scenario (TM): isolates and mildly refined ingredients are allowed in ingredients formulations and are selected based on the techno-functional properties

This means that formulations were either selected based on composition (scenarios 1 and 2) or on techno-functional properties (scenario 3). The composition and techno-functional properties were derived for both the ingredients and the products using the mathematical relations that describe these functional properties based on composition as described in **chapter 6 – mixed crops**. The composition is represented by the concentration (wt%) of protein, starch, fibre, and the combined rest of the components. The formulation space is the space in which the properties of the ingredients can be matched to the (predicted) product properties. The resulting formulation space revealed the window in which ingredients could be

optimised for their environmental impact, for example in terms of the lowest global warming potential, water use, or raw material use.

A portfolio of 14 food products was created (Figure 7.2). The composition of each of these products (e.g., a sandwich spread or coffee creamer) was derived from existing products on the market. Only the main macro components were considered. For that, two to five products from different brands were evaluated for each product and the composition was either an average of their compositions or at least well within the range of these products. The exact compositions can be found in the Appendix (Table A 7.1)

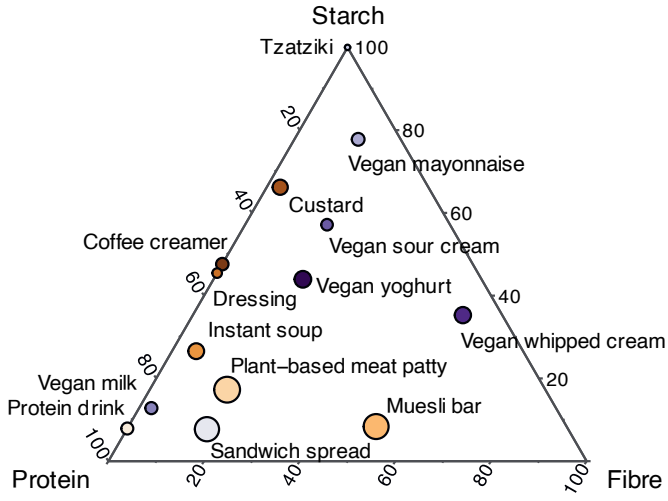

Figure 7.2. Dry base ratio between the protein, starch, and fibre content of selected food products. The size of the bubbles represents the relative dry base concentration of the ingredients replaced in each product. The function of the colours is to indicate the different products that are discussed in the text.

The part of the composition (e.g., protein or starch) of the products that was eligible to be replaced by the ingredients in this study was based on their ingredient lists. Mostly starch and protein isolates were identified. For example, if a starch isolate is added, the starch present in the product would be replaced. In some cases, only a part of one component had to be replaced: in the case of the instant soup, which consists of a lot of starch coming from a starch isolate but also from added potato

pieces, an estimated amount of starch isolate was replaced. As the rest fraction of these products consists of other ingredients (e.g., oils, creams, syrups, purees or acids) that are not considered in this study, the rest fraction was always set at 0 wt%. If a product does not require all components (e.g., no protein needs to be added), it was by default set to 0.5%. The reason for this is that otherwise no ingredient can be assigned to the products, as ingredients always contain multiple components. For each product, relevant key techno-functional properties were identified and modelled using the relationships developed in **chapter 6 – mixed crops** (Table 7.1). It was assumed that the ingredients in the products behave like the ingredients that were used in this study. For the viscosity of a protein drink, vegan milk (e.g., oat-based milk), tzatziki, and a vegan whipped cream the relationship that describes the unheated viscosity was used for these products. The reason for this is that these products are commonly not consumed at elevated temperatures. The impact of heating during processing for these examples was also considered minimal as the optimal solubility of all components should be maintained. The relation for the heated viscosity is used for heated products like coffee creamer, instant soup, and vanilla custard which are heated more extensively during the process. For the coffee creamer, vegan milk (to use as a barista version), and vegan whipped cream the foaming capacity is also considered. The products that were considered to include gel stiffness are the (soft) solid foods, such as a muesli bar, a plant-based meat patty (not a fibrous meat analogue), and a sandwich spread. Lastly, emulsion stability was considered important for products based on emulsions, which are the vegan mayonnaise and sour cream. It should be borne in mind that the above mentioned aspects are all assumptions for a model system to show the proof of concept yet are of course a lot more complex in reality. The formulation space to obtain the right techno-functional properties for each product was created in three steps based on the method described in **chapter 4 – the concept**. 1) We created a grid of compositions of protein, starch, fibre, and a rest fraction all varying from 0-25 wt% in Python 3.6 using the Constraint package 1.4.0 (Figure 7.3A). The emulsion stability, foam capacity, and unheated viscosity were based on models with a split in protein and fibre according to the origin. Therefore, another grid with split variables was created for these functional properties. As the formulation space is continuous,

Table 7.1 Summary of the techno-functional properties considered for each product. The number of possible formulations indicates the number of formulations from the original grid that results in the desired functional properties, as determined by the models in chapter 6 – mixed crops.

Product	Functional properties	Value	# possible formulations
Coffee creamer	Heated viscosity Foam overrun	133mPa.s 121%	999
Custard	Heated viscosity	508 mPa.s	1,781
Dressing	Emulsion stability	11 µm/s	13,273*
Instant soup	Heated viscosity	285 mPa.s	590
Muesli bar	Gel stiffness	127 kPa	8,152
Plant-based meat patty	Gel stiffness	88 kPa	8,630
Protein drink	Unheated viscosity	33 mPa.s	8,480*
Sandwich spread	Gel stiffness	26 kPa	5,474
Tzatziki	Heated viscosity	937 mPa.s	2,803
Vegan mayonnaise	Emulsion stability	48 µm/s	15,666
Vegan milk	Unheated viscosity Foam overrun	32 mPa.s 149 %	4,008*
Vegan sour cream	Emulsion stability	23 µm/s	95,813
Vegan whipped cream	Unheated viscosity Foam overrun	31 mPa.s 91 %	629*
Vegan yoghurt	Gel stiffness	5 kPa	1,220

*Formulations created with a grid size step of 1.3

the number of points to consider in principle is infinitely large. We used a grid step size of 0.6 or 1.3 wt% to limit the computing times to reasonable values but maximize the number of compositions. The first grid (protein, starch, fibre, rest) resulted in about 150,000 compositions with a step size of 0.6 wt%. With the split in protein and fibre, the grid resulted in about 10 million compositions. The latter was reduced for some products with a step size of 1.3 wt% to 170,000 compositions to reduce computing time. 2) The grids are subsequently translated to the techno-functional properties using the previously selected machine learning algorithms from **chapter 6 – mixed crops**, which is illustrated in Figure 7.3B for the foam overrun and unheated viscosity. Every point describes one composition of protein, starch, fibre, and rest fraction that results in the desired overrun and viscosity (illustrated in the squares for two products). Notably, compositions with multi-functional properties can now easily be identified.

There is a wide range of possible functional properties that can be obtained. Yet not all techno-functional properties calculated from the compositional grid are of interest at the same time; only the ones required by the products. The required functional properties of the products were calculated and presented in Table 7.1. In the last step 3), the formulation space; hence, compositions that result in the desired product properties can now be extracted. Every product now has a formulation space of possible formulations that result in the desired functional properties, as is shown for the vegan milk and whipped cream in Figure 7.3C. In our discretized space, this resulted in approximately 600 to 95,000 formulations per product (Table 7.1 and Appendix Figure A 7.1).

The formulation spaces were reduced to feasible formulation windows in which the formulations (combinations of ingredients) meet the techno-functional requirements for each product. Not all formulations (hence, compositions) in the formulation space could be created with the available ingredients, due to the constraints in the ingredients' composition. Therefore, a Linear Optimisation module (PuLP 2.6.0) was used to calculate the possible blends of ingredients that could match the desired compositions. These possible ingredient blends were then

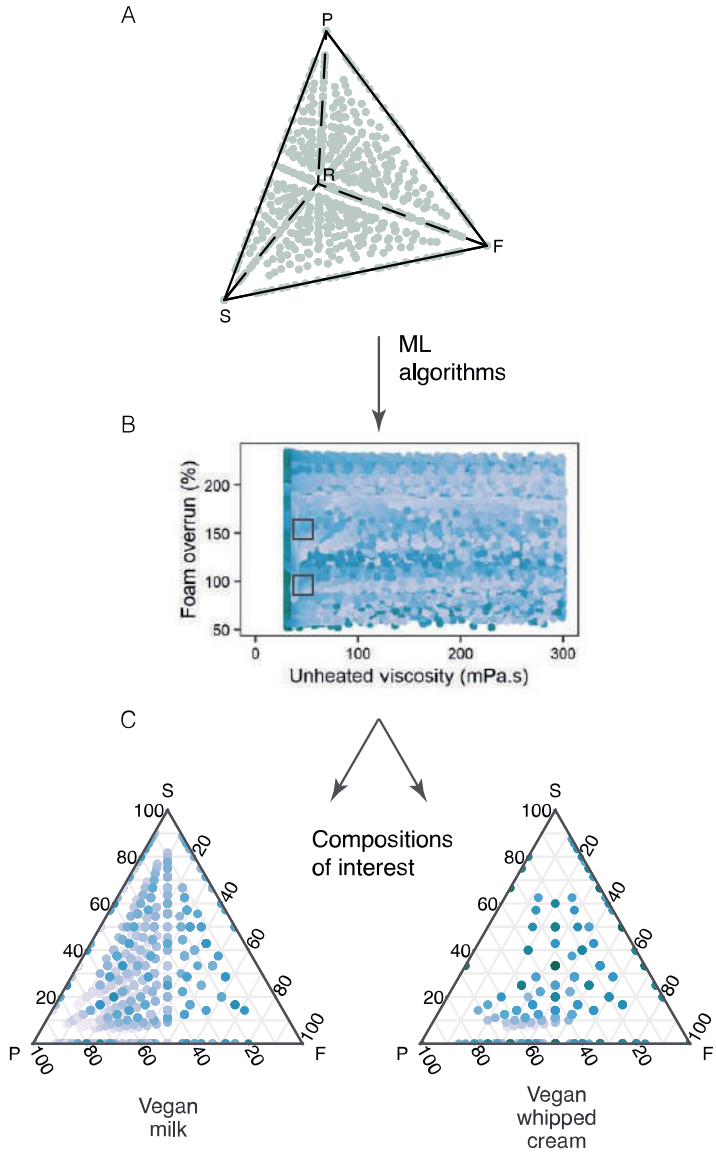

Figure 7.3 Creation of the formulation spaces with (A) the grid of compositions (Protein, Starch, Fibre, Rest) that is used to calculate the techno-functional properties for the complete grid here illustrate for foaming capacity and unheated viscosity (B). The compositions that results in the desired properties (in the squares) are now extracted here illustrated for the vegan milk and whipped cream (C). Similar illustrations for the other products can be found in Appendix Figure A 7.1. The colour intensity in B and C indicates the concentration of the formulations, with lighter colours higher concentrations.

minimised for one of three objective functions: global warming potential (GWP), water use, and the total amount of raw materials. For every formulation, the fraction (X) of ingredient (i) was fitted while minimizing these three objectives using $min \sum_{i=1}^{n} w_i \cdot X_i$. The values of the weights (w_i) differ per objective function and represent the specific 'embedded' global warming potential, the specific embedded water use, or the inverse of the yield of each ingredient (Appendix Table A 7.2).

To ensure that the ingredients that were blended together matched the composition from the formulation space, the blend was constrained by the composition that the scenario allows per formulation. For the base and CM scenario, only one formulation, or composition, is allowed, which is the composition of the product (Figure 7.2). In contrast, for the TM scenario, this is done for all possible formulations. In general, for all formulations, the requirements for each component (protein, starch, fibre, rest) were constrained by $\sum_{i=1}^{n} q_{i,c} \cdot X_i = q_{f,c}$, in which $q_{i,c}$ is the dry mass fraction (q) of each component (c) in each ingredient i (YPI, YFI, etc) multiplied with the fraction (X_i) of ingredient i in each formulation f. This has to be equal to the dry mass fraction q of composition c allowed in possible formulation f. These compositions were again based on the compositional match or generated from the formulation space.

The relative compositions of the formulations for the complete portfolio are presented in Figure 7.4 and separately per product in the Appendix (Figure A 7.2). In the TM scenario, there was a total of approximately 7,000 formulations per objective that could be matched to the products. As the aim of this study is to find the most sustainable ingredient formulations in the window of the possible formulations, only the solutions that score best in each objective function (i.e., lowest GWP, water usage, and raw materials) were extracted for each product and presented in this Figure. In scenarios 1 and 2, a feasible solution (formulation) was found for all products except for the vegan whipped cream. In the TM scenario, two products could not be matched, which were the coffee creamer and vegan mayonnaise. These products are excluded in all scenarios from the Figure and further analysed to ensure a fair comparison. The presence of unmatched products indicates that these compositions could not be matched by the available ingredients

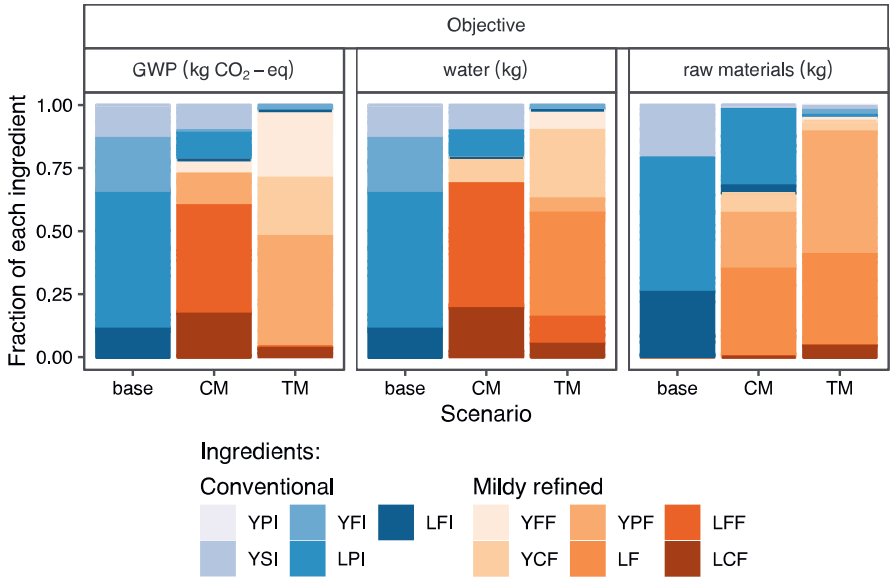

Figure 7.4 The fraction of each ingredient for the best solution for each objective (14 per objective) for the total product portfolio for the three objective functions (GWP, total ingredients, and water). The blue colours represent conventional ingredients, and the orange colours the mildly refined ones.

in this study. This is either a limitation of the composition of the ingredients or because some compositions are missed due to the finite grid step size of limitations of the functional properties.

The formulation in the base scenario only includes the conventionally fractionated ingredients for all objectives, as only these were allowed. The GWP and water as objective deliver ingredient formulations with the same composition, which means that the ingredients selected in these formulations have both the lowest GWP and water use. With the raw materials as the objective function, the ingredients with the highest yield are selected. The same applies to the CM scenario, but now mildly refined ingredients are also selected, since all of these have a relatively low GWP and water use, and the unrefined flours from yellow pea and lupine of course have a higher yield (100%). The study by Jonkman et al.[18] also found that mildly refined ingredients could be matched based on composition, and with that reduce the resource use. In the TM scenario, the share of mildly refined ingredients is even

larger for each objective. Yellow pea ingredients are favoured with the GWP objective, whereas lupine ingredients are favoured for the objective function using water usage and raw materials. Both can be explained by their relative impacts.

The cumulative impacts for each product in the portfolio for the three different scenarios and objective functions are shown in Figure 7.5. The impact (obtained in **chapter 3 - sustainability**) of each formulation was calculated per ton (1000 kg) of ingredient formulation, which consists of the selected dry ingredients using: $impact = \frac{\sum_{i=1}^{n} X_i impact_i}{\sum_{i=1}^{n} X_i} \cdot 1000\ kg$. The total raw material usage for each formulation needed to produce 1 ton was also calculated using the ingredient yield with: $raw\ material = \frac{\sum_{i=1}^{n} \frac{X_i}{yield_i}}{\sum_{i=1}^{n} X_i} \cdot 1000\ kg$.

The impact of the total product portfolio in the TM scenario is lower compared to the base and the CM scenarios. The GWP of the total portfolio is reduced by almost 50% from the impact generated by a selection based on composition with mildly refined ingredients, and by more than 70% compared to the base scenario (using only isolates). The water use is reduced even more, which is attributed to the use of mainly dry fractionated ingredients in these formulations. This means that the water usage is almost completely from the cultivation of the crop, which is relatively low. Notably, even when one objective of the TM scenario was optimized (e.g., GWP), the other objectives (water use, raw material use) were either similar or even lower than for the base and CM scenarios. For example, while minimizing the GWP of the product in the TM scenario, the water and raw material use was still lower than in the base and CM scenarios.

Interestingly, the number of raw materials required was lower in the CM scenario compared to the base scenario. This is surprising as the underlying hypothesis of this thesis was that with a selection based on composition, more raw materials are needed when incorporating milder refined ingredients due to their high fibre and other carbohydrate-rich ingredients, as was found by Jonkman et al.[18] One main difference is that in that study, the impact allocated to the ingredients did not include the impact of cultivating the raw materials. Apparently, if this is included, as is done

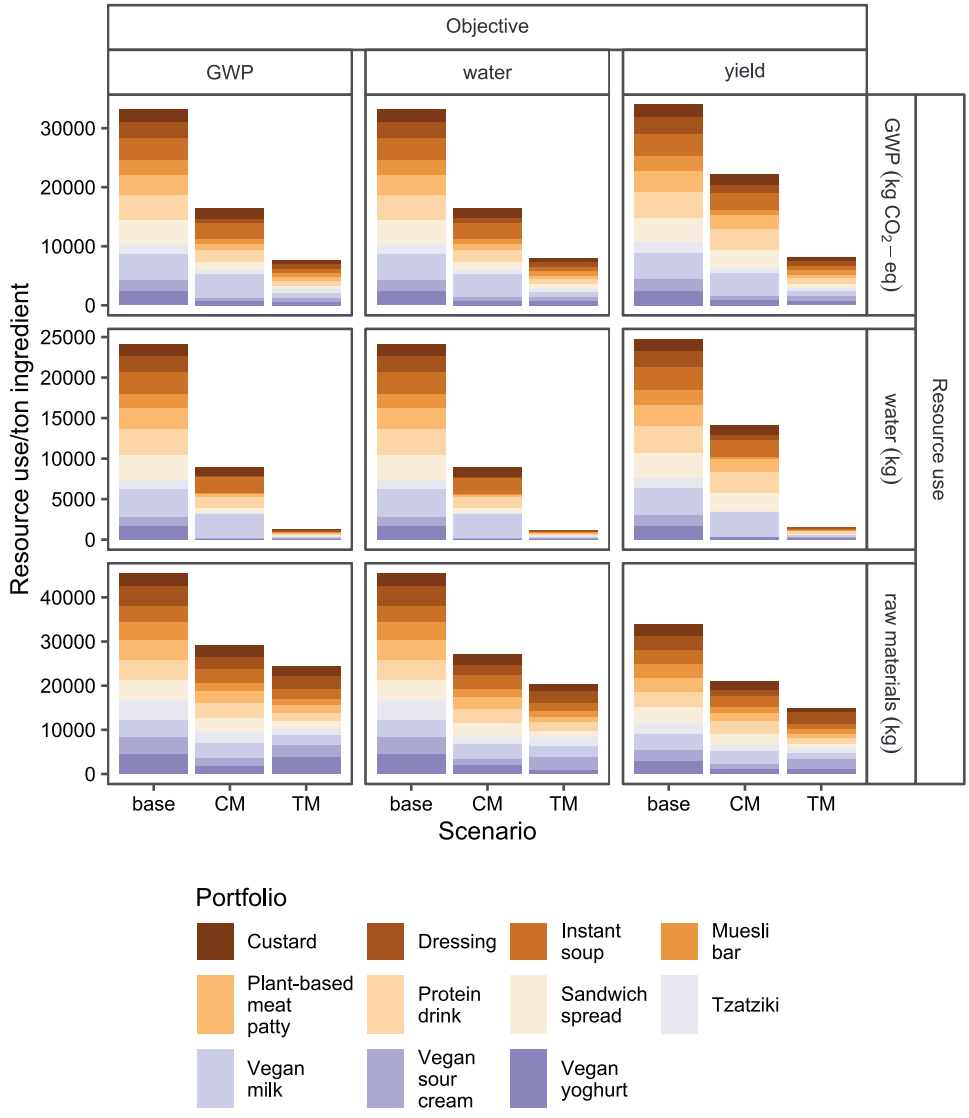

Figure 7.5 Impact of the product portfolio for three different objectives GWP (kg CO_2-eq /ton ingredient per product), water usage (kg water/ton ingredient per product) and total raw materials (kg/ton ingredient per product).

in this study, the ingredients assigned to the formulations will be used more efficiently to both minimize the impact of processing the crops into ingredients, but also minimize the impact of the cultivation. This evidently results in the use of fewer raw materials. In this study, the reuse of the co-production of other ingredients (e.g. starch isolate during the production of a protein isolate) for other products in the portfolio is not even considered yet, which could result in even lower raw material usage.

Jonkman et al. [18] also provided estimated costs to process a ton of raw materials through either the conventional or dry pathways. These estimations were used to calculate the costs of the portfolio presented in this study in two ways: 1) based on the amount of ingredients that was assigned to the products or 2) based on the total raw materials that were required (Figure 7.5). Both methods show that to produce a ton of ingredients for the complete portfolio, the costs decrease substantially with the use of milder refined ingredients in the CM scenario and even further in the TM scenario (Figure 7.6). The decrease in costs is attributed to the larger share of cheaper milder refined ingredients used and the lower number of raw materials needed to produce the ingredients for the portfolio.

Notably, the number of raw materials required for the dressing and vegan sour cream with the TM is larger than with the CM (Appendix Figure A 7.3). This means that there is a composition with a lower impact that is not present in the revealed formulations space. This suggests that the techno-functionality is probably not reached with the composition obtained from the CM scenario. This was indeed the case for the dressing, as the emulsion stability from the composition in the CM scenario is twice as large as the desired stability in the TM scenario. For the vegan sour cream, the desired emulsion stability is actually achieved with the formulation form the CM scenario. This indicates that the composition of the CM scenario was not included in the compositional grid. This could be solved by decreasing the step size.

The results indicate that the TM is for most products valuable in terms of environmental impact, costs, and techno-functional properties. However, as some products cannot be matched yet, the optimization of these products can still be

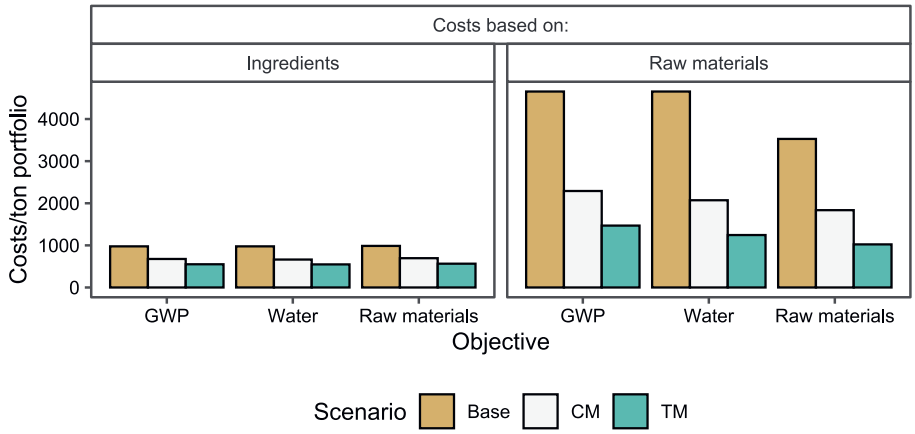

Figure 7.6 Estimated costs of the total product portfolio, with 95.82 and 85.02 €/ton ingredients for conventional lupine and yellow pea ingredients respectively and 48.48 €/ton ingredients for both dry fractionated yellow pea and lupine. The costs are shown for each objective and scenario calculated with two methods: based on ingredients used and total raw materials. Data of costs are obtained from another study.[18]

improved. In such a situation one may use an experimental optimization procedure in which the models are used to direct and select formulation areas of interest. Repeating this should result in finding the best possible formulation for that product.

7.4 Why is a techno-functional selection more resource use efficient?

Figure 7.4 depicts the ingredients that the portfolio is composed of for each scenario. Even for the CM scenario, mildly refined ingredients were selected instead of conventional ingredients since the used objective was to minimize the global warming potential, water use, and total ingredients usage in the formulations. In the TM scenario, the fraction of milder refined ingredients was larger, resulting in a lower impact for most products (Figure 7.5). The total global warming potential of the portfolio could be reduced by 50% and for water even more, relative to the CM scenario, and more than 70% relative to the base scenario. The latter means that compared to the current way products are formulated using conventional ingredients, a 70% reduction of GWP can be achieved with the use of milder refined

ingredients and a selection based on techno-functional properties in a complete product portfolio.

One reason for this lower environmental impact is that the match based on the functional properties enlarges the formulation window. More specifically, in the CM scenario, there is only one composition (with one of the products) that can match the ingredients. In the discretized techno-functional formulation space created in this study, this was increased to 500-95,000 discretized formulations per product, which of course will be much more when using smaller grid steps. It is also hypothesized that products that are still unmatched in the TM scenario will yield a similar or better formulation than the CM scenario with smaller grid steps. This can only mean that the total impact of the TM scenario will be lower. These formulations represent a larger window in the composition matrix as is depicted in Figure 7.3C for a few products, and in Appendix Figure A 7.1 for all products. Here one can see that differences in composition can be matched with differences in overall concentrations (dry weight), to end up with the right properties. The larger window implies more possibilities for the application of milder refined ingredients and hence a larger probability of finding a solution with a lower environmental impact.

In addition, mildly refined ingredients deliver also more functionality per unit of environmental impact compared to conventionally produced ingredients, as illustrated for GWP in Figure 7.7. In this figure, each functionality of each ingredient (ranging from 1-22 wt%) is expressed per kg CO_2-eq emissions. Isolates supply less functionality per kg CO_2-eq (GWP) compared to milder refined ingredients. The exception is the emulsion separation velocity, in which milder refined ingredients with a similar GWP are less stable (higher separation velocity) compared to isolates. Geerts et al.[8] showed in a similar analysis that mildly refined ingredients also contribute more viscosity per unit of exergy efficiency.

The conclusion is that a selection based on techno-functional ingredients is more resource use efficient since the formulation space is increased. As a result, the incorporation of milder refined ingredients is facilitated, which are generally – but not always – intrinsically more sustainable when considering their techno-functional role in foods.

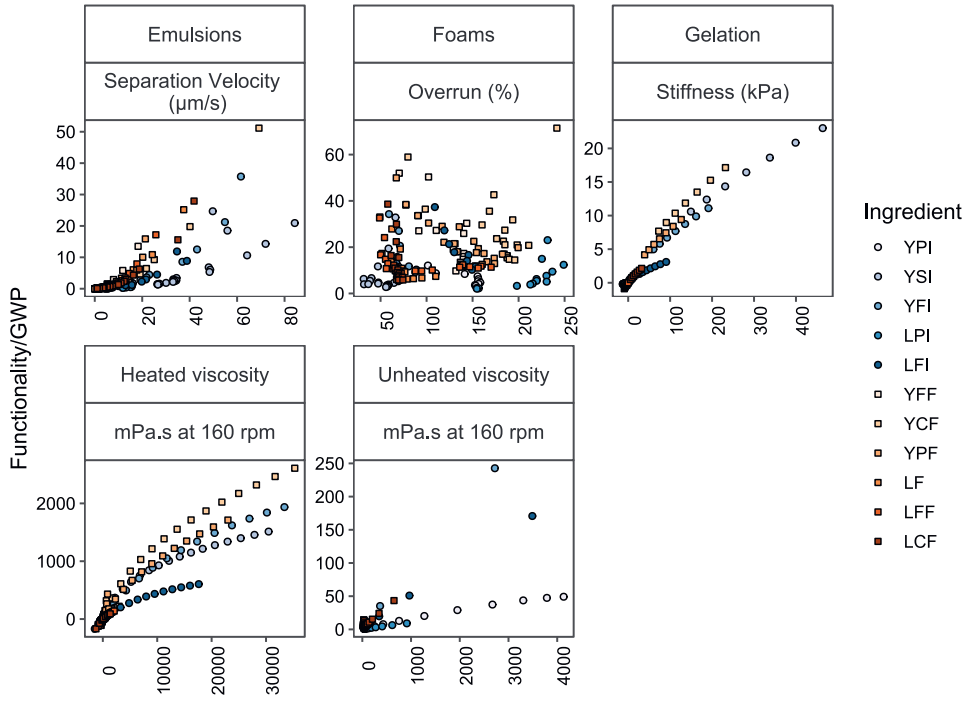

Figure 7.7 Functionality expressed per unit global warming potential (kg CO_2-eq) as a function of each functional property. The ■ indicate mildly refined ingredients and the ● indicate conventional highly refined ingredients.

7.5 Data collection

The modelling of techno-functional properties in **chapters 4, 5, and 6** required the collection of a large dataset, which is time and resource intensive. The datasets in this thesis were developed incrementally over four years by manual data collection. This may be undesired. Ultimately, the data collection could be incorporated into product development, where results from trial and error experiments are stored in structured databases that are easily accessible. When properly managed with experimental designs that take the existing data into account, the accumulation of observations will over time lead to sufficient data to model properties. This can be done for new, still unmeasured functionalities, but also for the expansion of existing datasets, to incorporate for example new processing conditions or to improve the accuracies of properties that require more observations. The latter is relevant for the

models for foams and emulsions as more observations can lead to higher accuracy, as was shown in **chapter 6 – mixed crops**. In addition, many studies do similar analyses on for example viscosity or foaming with similar ingredients of the same origin. Standardization of protocols to measure functional properties would allow sharing of data on functional properties. This was for example possible in the study presented in the introduction, in which many recipes from chocolate cookies were extracted from the internet and then modelled using hybrid mechanistic and machine learning models.[21] Automatic sampling and experimentally searching for an optimal formulation could further speed up the data collection, as was performed for the viscosity and turbidity of whey protein, using Gaussian processes to move to the Pareto front.[194] The work in this thesis indicates that an overall model (**chapter 6 – mixed crops**) is important to find the areas of interest prior to experimental optimization using automatic sampling. This will avoid long experimental series migrating towards the right formulation area, and ending up in local minima, which in fact are probable, given the inherent nonlinearity in the system.

7.6 Formulation validation

One relevant next step is the validation of the formulations presented in this study. Although this was out of scope for the current study, it is important to experimentally validate the formulations that were identified as optimal, and thus test whether the predictions of the models are reliable. The product formulations must be prepared and tested on the techno-functional properties. The models for the individual technical properties in this study were created with mixtures of the different ingredients and were always tested based on an independent dataset. It is therefore expected that the accuracy of the presented formulations does not deviate much from the current accuracy.

7.7 How to improve and extend current models?

The current approach can be used to pinpoint relevant formulation areas for milder refined ingredients, which can replace time-intensive trial and error experiments. However, as we have seen in chapter 6 – mixed crops, the test error is in some cases substantially higher than the training error. In addition, the behaviour on foaming and emulsion is modelled with a random forest algorithm, which results in some

cases in artefacts in their behaviour. Therefore, it can be concluded that the prediction accuracy of some relationships can still be improved.

In this work the training and test samples were chosen manually. The accuracy could benefit from a standardized approach, using for example a Kennard-Stone algorithm to always ensure an equal spread within the composition matrix. This can also be applied to an existing dataset to identify new samples to be measured. Besides the selection of training and test samples, the used methods can also be improved in terms of reproducibility. For example, the measurement of the foam capacity is to some degree subjective, as the foam is not completely flat, and thus the foam height does not have one definite value. Lastly, in this study the composition was assumed sufficient to describe the functional properties. Splitting the proteins according to their processing history did not make any significant difference. Besides composition, the structure of the ingredients themselves and of other ingredients present in the formulation will also influence the final properties. For example, an emulsion consists of 10wt% oil, and this oil will definitely influence the viscosity. Another example is that the particle size of the flours influences bubble formation during foaming.

A possible lead to further improve the approach would be to create an intermediate layer of physical properties that would have a better, more generic relation with the ultimate technical functionalities. If functional properties can be described by their physical properties such as the water holding capacity or particle size, the formulation of products can become more generic as opposed to the conversion to composition and back. For example, the water competition between different phases that relates to the pasting properties of yellow peas could be quantified using a polymer blending law.[174] This model provides information that can be implemented and combined with the current models to create so-called hybrid models.

As was mentioned before, the techno-functional properties selected in this study work were chosen for a demonstration of the concept. Interactions with other ingredients and conditions (e.g., salt or pH, temperature) will have to be included. The presented approach will still be appropriate, as machine learning algorithms are suitable to model systems with many input variables. Refinement of the models

could and should be done by incorporating the effects of other components, such as electrolytes. Since these effects are relatively well understood, one could imagine here that the models could be combined with mechanistic models to describe these effects. With an optimized data collection and/or automated method (e.g., incorporation of Kennard Stone sampling), the amount of data collected can be minimized. In addition, other functional properties can be added. Of course, there are many more techno-functional properties (e.g., colour, degree of shear thinning, transparency, etc), but one should also consider including aspects of the sensory perception by consumers (e.g., taste and colour), and the bioavailability and nutritional functionalities of the foods.

7.8 A final word

The effect of a functionality-driven selection of ingredients as opposed to a compositional selection on the environmental impact of food formulations was assessed. The selection was shown for matching a portfolio of 14 products based on the compositions or the techno-functional properties (gelation, viscosity, emulsion stability, foaming capacity). These techno-functional properties could be quantified using machine learning. The obtained relationships were used to identify the formulations space that results in the desired techno-functional properties.

The impact of the total product portfolio was already reduced by 50% with the use of mildly refined ingredients next to conventional isolates when matching ingredients to products based on compositions. Matching based on techno-functional properties substantially enlarged the formulation space for all products. The global warming potential of the total product portfolio with a techno-functional match could be reduced by 70% compared to the composition-based selection with only isolates. The total water usage was reduced even more. In addition, the milder refined ingredients contribute more functionality (e.g., viscosity) per kg CO_2-eq in most cases. The reduced impact with a selection based on techno-functional properties is attributed to the higher intrinsic functionality of mildly refined ingredients intrinsically per unit of global warming potential and that with a functionality-driven formulation, the space in which these ingredients can be matched to the final products is enlarged.

This thesis aimed to assess to what extent the use of a functionality-driven selection of ingredients using machine learning may lead to more sustainable food production. Indeed, it was found that machine learning is key to translating the composition to techno-functional properties. Food product formulations can be created that are optimized for resource use and composition. By revealing the whole formulation space, local minima are prevented. Yet, the current assessment is only meant as a first step and a demonstration of the concept. The formulations at this moment only give (so far) limited accuracy. In a next step, the exact formulations can be further narrowed down using targeted experiments.

7.9 Appendix

Table A 7.1 Product portfolio with compositions, functional properties considered, and ingredients replaced.

Product	Property 1	Property 2	Protein (g/100g dry matter)	Starch (g/100g dry matter)	Fibre (g/100g dry matter)	Rest (g/100g dry matter)	Ingredient replaced 1	Ingredient replaced 2
Coffee creamer	Heated viscosity	Foam	2.15	2.00	0.00	0	Protein isolate	Part of the starch
Custard	Heated Viscosity		2.20	4.70	0.20	0	Protein isolate	Starch isolate
Dressing	Emulsion		1.20	1.00	0.00	0	Protein isolate	Gums
Instant soup	Heated Viscosity		5.40	2.11	0.44	0	Protein isolate	Starch isolate
Muesli bar	Gelation		10.00	2.14	13.00	0	Fibre isolate	Starch isolate
Plant-based meat patty	Gelation		18.13	4.75	4.42	0	Protein isolate	Starch isolate
Protein drink	Unheated viscosity		3.50	0.27	0.00	0	Protein isolate	
Sandwich spread	Gelation		17.61	1.78	3.87	0	Protein isolate	Fibre isolate
Tzatziki	Heated Viscosity		0.00	1.10	0.00	0	Starch isolate	
Vegan mayonnaise	emulsions		0.40	3.53	0.60	0	Protein isolate	Starch isolate
Vegan milk	Unheated viscosity	Foam	3.32	0.47	0.13	0	Protein isolate	Gums
Vegan sour cream	Emulsion		0.90	2.01	0.60	0	Starch isolate	Protein isolate
Vegan whipped cream	Unheated viscosity	Foam	0.70	3.00	4.80	0	Protein isolate	Inulin
Vegan yoghurt	Gelation		3.37	4.00	1.72	0	Protein isolate	Starch isolate

Table A 7.2 Environmental impact of each ingredient. The global warming potential in kg CO_2-eq and blue water use are expressed per kg of ingredient. Data is obtained from chapter 3 – sustainability.

	kg CO_2-eq/kg	Water/kg	Yield
YPI	4.20	0.08	0.22
YSI	1.01	0.03	0.44
YFI	0.86	0.03	0.14
YFF	0.68	0.02	0.23
YCF	0.68	0.02	0.77
YPF	0.67	0.02	1.00
LPI	5.02	0.05	0.27
LFI	1.46	0.03	0.37
LF	0.74	0.002	1.00
LFF	0.75	0.002	0.33
LCF	0.75	0.002	0.67

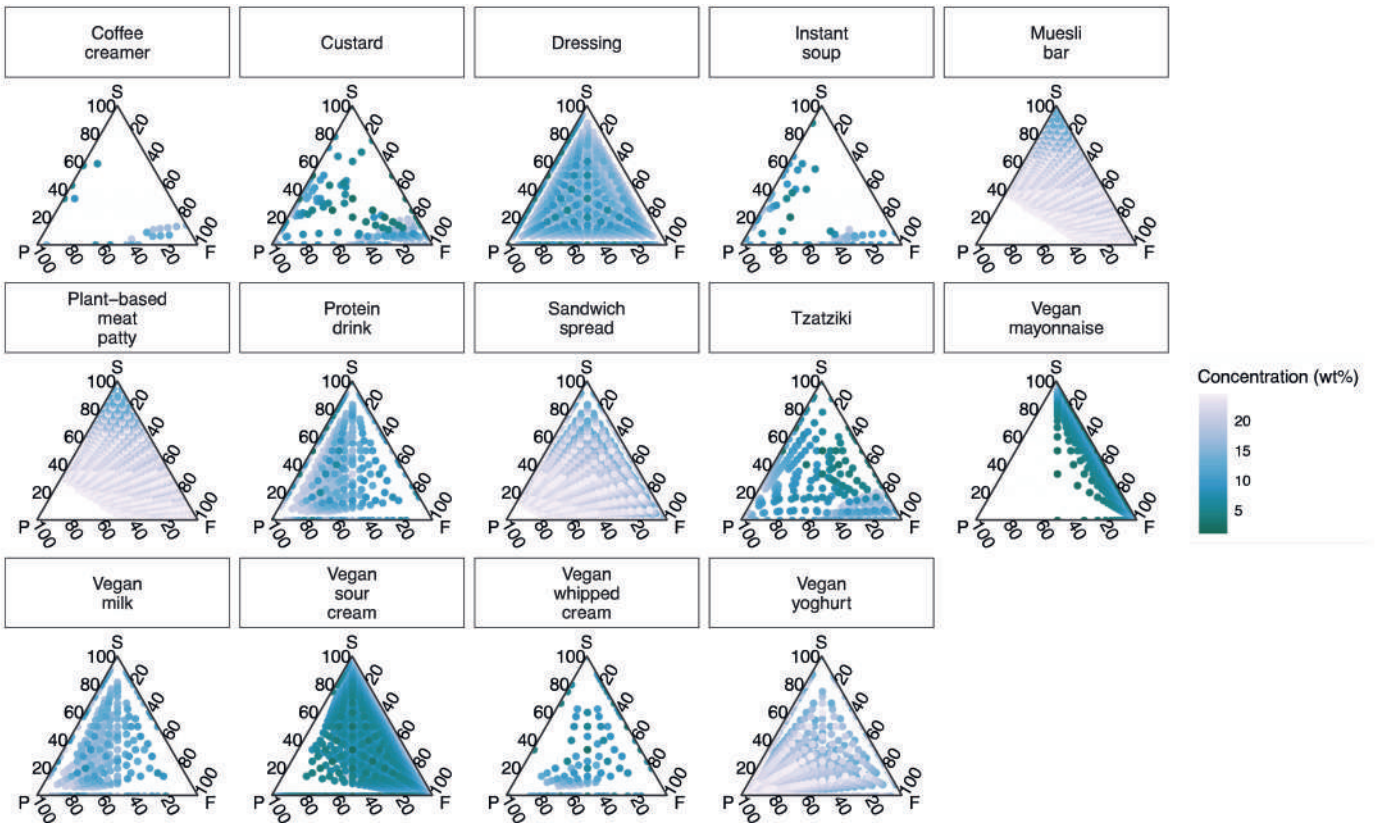

Figure A 7.1 Formulation spaces for all products from the portfolio.

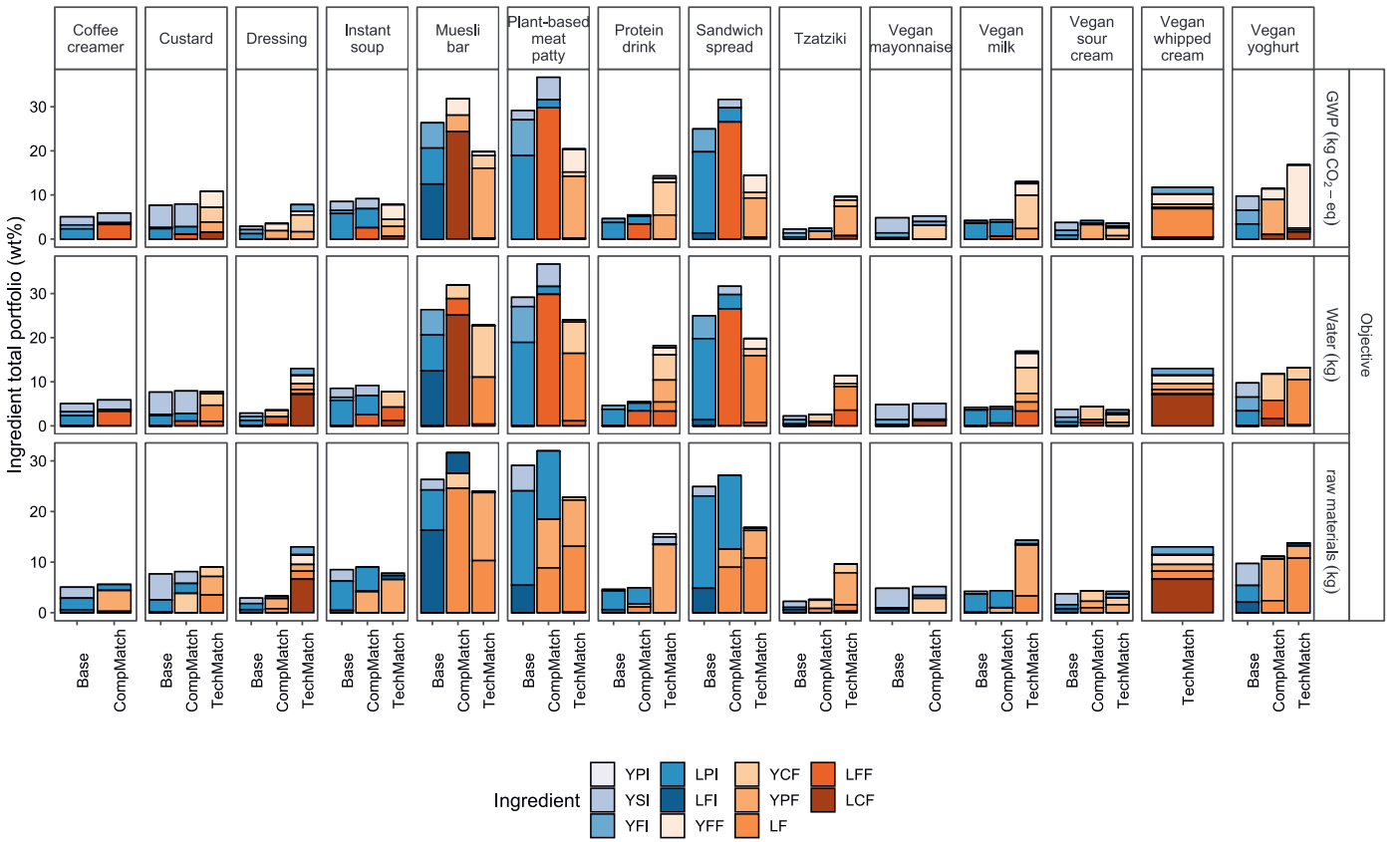

Figure A 7.2 Overview of the composition of the ingredient formulations for each product and objective function.

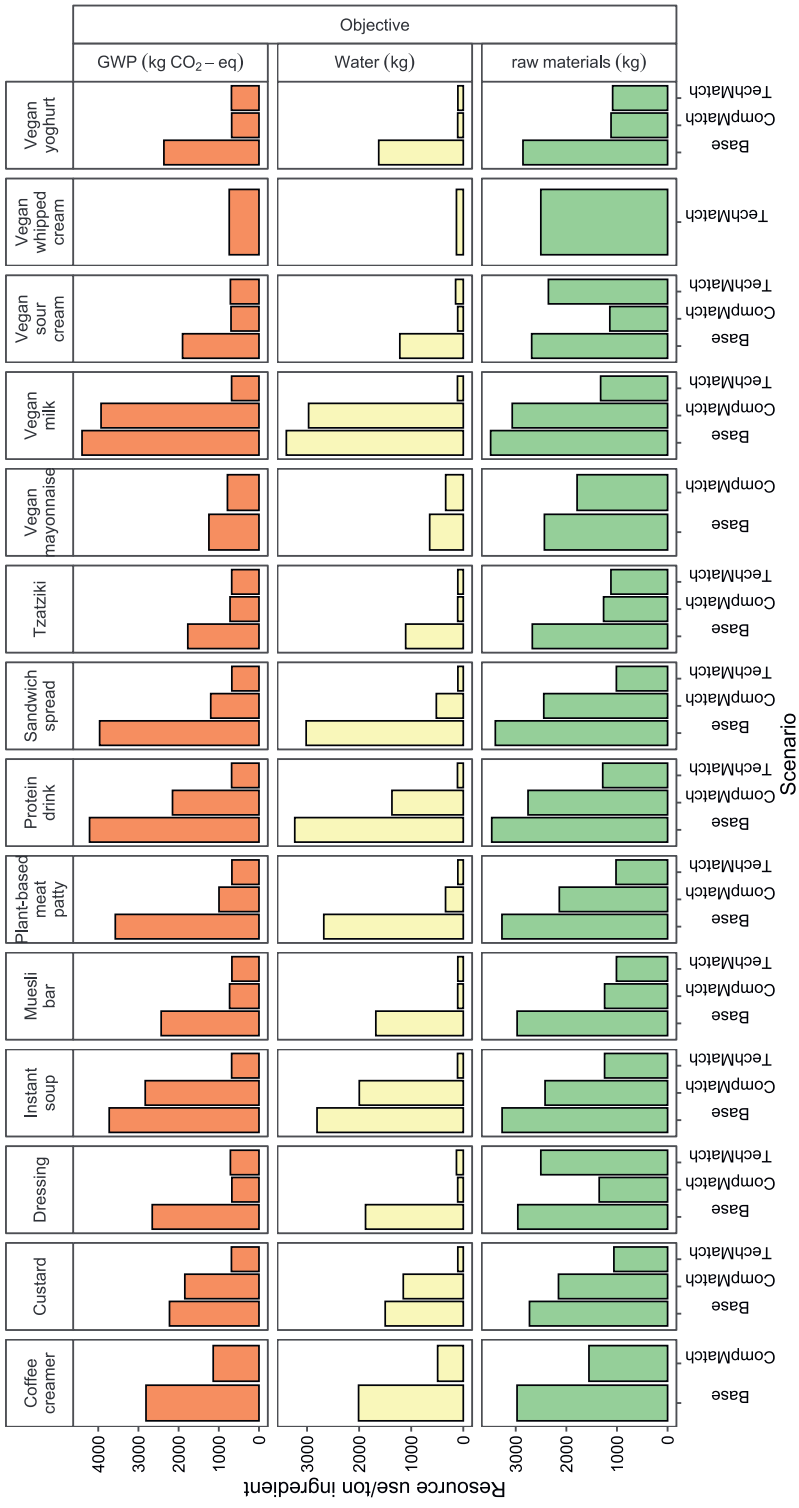

Figure A 7.3 Overview of the impact expressed in resource use (GWP, water, or raw materials) of the ingredient formulations for each product and objective function.

References

1. van Boekel M, Fogliano V, Pellegrini N, Stanton C, Scholz G, Lalljie S, et al. A review on the beneficial aspects of food processing. Mol Nutr Food Res. 2010;54(9):1215–47.

2. van der Goot AJ, Pelgrom PJM, Berghout JAM, Geerts MEJ, Jankowiak L, Hardt NA, et al. Concepts for further sustainable production of foods. J Food Eng. 2016;168:42–51.

3. Apaiah RK, Linnemann AR, van der Kooi HJ. Exergy analysis: A tool to study the sustainability of food supply chains. Int Food Res J. 2006;39(1):1–11.

4. Anderson JW, Baird P, Davis RH, Ferreri S, Knudtson M, Koraym A, et al. Health benefits of dietary fiber. Nutr Rev. 2009;67(4):188–205.

5. Slavin JL. Position of the American Dietetic Association: health implications of dietary fiber. J Am Diet Assoc. 2008;108(10):1716–31.

6. Rodríguez-Ambriz SL, Martínez-Ayala AL, Millán F, Dávila-Ortíz G. Composition and functional properties of Lupinus campestris protein isolates. Plant Foods Hum Nutr. 2005;60(3):99–107.

7. Swamylingappa B, Srinivas H. Preparation and Properties of Protein Isolate from Hexane-Acetic Acid Treated Commercial Soybean Meal. J Agric Food Chem. 1994;42(12):2907–11.

8. Geerts MEJ, van Veghel A, Zisopoulos FK, van der Padt A, van der Goot AJ. Exergetic comparison of three different processing routes for yellow pea (Pisum sativum): Functionality as a driver in sustainable process design. J Clean Prod. 2018;183:979–87.

9. Lie-Piang A, Braconi N, Boom RM, van der Padt A. Less refined ingredients have lower environmental impact – A life cycle assessment of protein-rich ingredients from oil- and starch-bearing crops. J Clean Prod. 2021;292.

10. Tanger C, Engel J, Kulozik U. Influence of extraction conditions on the conformational alteration of pea protein extracted from pea flour. Food Hydrocoll. 2020;107(April):105949.

11. Kornet C, Venema P, Nijsse J, van der Linden E, van der Goot AJ, Meinders M. Yellow pea aqueous fractionation increases the specific volume fraction and viscosity of its dispersions. Food Hydrocoll. 2020;99(June 2019):105332.

12. Maaroufi C, Melcion JP, de Monredon F, Giboulot B, Guibert D, le Guen MP. Fractionation of pea flour with pilot scale sieving. I. Physical and chemical characteristics of pea seed fractions. Anim Feed Sci Technol. 2000;85(1–2):61–78.

13. Schutyser MAI, van der Goot AJ. The potential of dry fractionation processes for sustainable plant protein production. Trends Food Sci Technol. 2011;22(4):154–64.

14. Zhu HG, Tang HQ, Cheng YQ, Li ZG, Tong LT. Electrostatic separation technology for obtaining plant protein concentrates: A review. Trends Food Sci Technol. 2021;113(June 2020):66–76.

15. Cai R, Klamczynska B, Baik BK. Preparation of bean curds from protein fractions of six legumes. J Agric Food Chem. 2001;49(6):3068–73.

16. Rivera del Rio A, Möller AC, Boom RM, Janssen AEM. In vitro gastro-small intestinal digestion of conventional and mildly processed pea protein ingredients. Food Chem. 2022;387(February).

17. Berghout JAM, Pelgrom PJM, Schutyser MAI, Boom RM, van der Goot AJ. Sustainability assessment of oilseed fractionation processes: A case study on lupin seeds. J Food Eng. 2015;150:117–24.

18. Jonkman J, Castiglioni A, Akkerman R, van der Padt A. Improving resource efficiency in the food industry by using non-conventional intermediate products. J Food Eng. 2020;287(May).

19. Taifouris M, Martín M, Martínez A, Esquejo N. Challenges in the design of formulated products: multiscale process and product design. Vol. 27, Curr. Opin. Chem. Eng. . Elsevier Ltd; 2020. p. 1–9.

20. Manzocco L, Nicoli MC. Food design: From the methodological approach to the case study of low-calorie syrups. Trends Food Sci Technol. 2002;13(12):422–9.

21. Zhang X, Zhou T, Zhang L, Fung KY, Ng KM. Food Product Design: A Hybrid Machine Learning and Mechanistic Modeling Approach. Ind Eng Chem Res. 2019 Sep 11;58(36):16743–52.

22. Mitchell T, Buchanan B, Dejong G, Dietterich T, Rosenbloom P, Waibel A. Machine Learning. Annu Rev Compu Sci [Internet]. 1990;4:417–33. Available from: www.annualreviews.org

23. Dixon MF, Halperin I, Bilokon P. Machine learning in finance: From theory to practice. Machine Learning in Finance: From Theory to Practice. Springer International Publishing; 2020. 1–548 p.

24. Jordan M, Mitchell T. Machine learning: Trends, perspectives, and prospects. Science (1979). 2015 Jul 17;349(6245):253–5.

25. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Vol. 23, Nature Reviews Molecular Cell Biology. Nature Research; 2022. p. 40–55.

26. Garre A, Ruiz MC, Hontoria E. Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty. Oper Res Perspect. 2020;7(February):100147.

27. Frizzarin M, O'Callaghan TF, Murphy TB, Hennessy D, Casa A. Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. J Dairy Sci. 2021;104(12):12394–402.

28. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R Second Edition. 2021.

29. Martinez-Castillo C, Astray G, Mejuto JC, Simal-Gandara J. Random Forest, Artificial Neural Network, and Support Vector Machine Models for Honey Classification. eFood. 2020 Feb;1(1):69–76.

30. Yu H, Liu J, Chen C, Heidari AA, Zhang Q, Chen H, et al. Corn Leaf Diseases Diagnosis Based on K-Means Clustering and Deep Learning. IEEE Access. 2021;9:143824–35.

31. Noordraven LEC, Petersen MA, van Loey AM, Bredie WLP. Flavour stability of sterilised chickpeas stored in pouches. Curr Res Food Sci. 2021 Jan 1;4:773–83.

32. Ying X. An Overview of Overfitting and its Solutions. J Phys Conf Ser. 2019;1168(2).

33. Aiking H. Future protein supply. Trends Food Sci Technol. 2011;22(2):112–20.

34. Xing Q, Dekker S, Kyriakopoulou K, Boom RM, Smid EJ, Schutyser MAI. Enhanced nutritional value of chickpea protein concentrate by dry separation and solid state fermentation. IFSET. 2020;59(November 2019):102269.

35. Ayerdi GA, Larbi R. Effects of refining process on sunflower oil minor components : a review. 23(2). OCL. 2016;23(2).

36. Gueguen J. Legume seed protein extraction, processing, and end product characteristics. Qual Plant Plant Foods Hum Nutr. 1983;32(3–4):267–303.

37. Politiek RGA, Bruins ME, Keppler JK, Schutyser MAI. Effect of oil content on pin-milling of soybean. J Food Eng. 2022;334(May):111149.

38. Srinivasan R, Singh V. Pericarp fiber separation from corn flour using sieving and air classification. Cereal Chem. 2008;85(1):27–30.

39. Aguilera JM, Crisafulli EB, Lusas EW, Uebersax MA, Zabik ME. Air Classification and Extrusion of Navy Bean Fractions. J Food Sci. 1984;49(2):543–6.

40. Tyler RT, Youngs CG, Sosulski FW. Air Classification of Legumes. I. Separation Efficiency, Yield, and Composition of the Starch and Protein Fractions. Vol. 58, Cereal Chem. 1981. p. 144–8.

41. King RD, Dietz HM. Air classification of rapeseed meal. Cereal Chem. 1987;64(6):411–3.

42. Vasanthan T, Bhatty R. Starch purification after pin milling and air classification of waxy, normal, and high amylose barleys. Cereal Chem. 1995;

43. Wu VY, Stringfellow AC, Inglett GE. Protein- and Beta-glucan enriched fractions from high-protein, high beta-glucan barley by sieving and air classification. Cereal Chem. 1994;71(3):220–3.

44. Wu VY, Stringfellow A. Enriched protein- and β-glucan fractions from high-protein oats by air classification. Cereal Chem. 1995;72(1):132–4.

45. Ferrari B, Finocchiaro F, Stanca AM, Gianinetti A. Optimization of air classification for the production of β-glucan-enriched barley flours. J Cereal Sci. 2009;50(2):152–8.

46. Wu VY, Doehlert DC. Enrichment of β-glucan in oat bran by fine grinding and air classification. LWT. 2002;35(1):30–3.

47. Challa R, Srinivasan R, To F. Fractionation of soybean meal, cottonseed meal and wheat middlings using combination of sieving and air classification. Anim Feed Sci Technol. 2010;159(1–2):72–8.

48. Pelgrom PJM, Berghout JAM, van der Goot AJ, Boom RM, Schutyser MAI. Preparation of functional lupine protein fractions by dry separation. LWT. 2014;59(2P1):680–8.

49. Pelgrom PJM, Vissers AM, Boom RM, Schutyser MAI. Dry fractionation for production of functional pea protein concentrates. Int Food Res J. 2013;53(1):232–9.

50. Wu VY, Nichols NN. Fine grinding and air classification of field pea. Cereal Chem. 2005;82(3):341–4.

51. Pelgrom PJM, Boom RM, Schutyser MAI. Method Development to Increase Protein Enrichment During Dry Fractionation of Starch-Rich Legumes. Food Bioproc Tech. 2015;8(7):1495–502.

52. Schutyser MAI, Pelgrom PJM, van der Goot AJ, Boom RM. Dry fractionation for sustainable production of functional legume protein concentrates. Trends Food Sci Technol. 2015;45(2):327–35.

53. Barakat A, Mayer C. Electrostatic Separation as an Entry into Environmentally Eco-Friendly Dry Biorefining of Plant Materials. J Chem Eng Process Technol. 2017;08(04):4–10.

54. Wang J, Zhao J, de Wit M, Boom RM, Schutyser MAI. Lupine protein enrichment by milling and electrostatic separation. IFSET. 2016;33:596–602.

55. Xing Q, de Wit M, Kyriakopoulou K, Boom RM, Schutyser MAI. Protein enrichment of defatted soybean flour by fine milling and electrostatic separation. IFSET. 2018;50(July):42–9.

56. Tabtabaei S, Jafari M, Rajabzadeh AR, Legge RL. Solvent-free production of protein-enriched fractions from navy bean flour using a triboelectrification-based approach. J Food Eng. 2016;174:21–8.

57. Basset C, Kedidi S, Barakat A. Chemical- and Solvent-Free Mechanophysical Fractionation of Biomass Induced by Tribo-Electrostatic Charging: Separation of Proteins and Lignin. ACS Sustain Chem Eng. 2016;4(8):4166–73.

58. Sibakov J, Abecassis J, Barron C, Poutanen K. Electrostatic separation combined with ultra-fine grinding to produce β-glucan enriched ingredients from oat bran. IFSET. 2014;26:445–55.

59. Hemery Y, Holopainen U, Lampi AM, Lehtinen P, Nurmi T, Piironen V, et al. Potential of dry fractionation of wheat bran for the development of food ingredients, part II: Electrostatic separation of particles. J Cereal Sci. 2011;53(1):9–18.

60. Bittner JD, Hrach FJ, Gasiorowski SA, Canellopoulus LA, Guicherd H. Triboelectric belt separator for beneficiation of fine minerals. Procedia Eng. 2014;83:122–9.

61. Xing Q, Utami DP, Demattey MB, Kyriakopoulou K, de Wit M, Boom RM, et al. A two-step air classification and electrostatic separation process for protein enrichment of starch-containing legumes. IFSET. 2020;66(August):102480.

62. Zhu HG, Tang HQ, Cheng YQ, Li ZG, Tong LT. Novel electromagnetic separation technology for the production of pea protein concentrate. IFSET. 2021;70(March):102668.

63. Plant AR, Moore KG. The protein, lipid and carbohydrate composition of protein bodies from Lupinus angustifolius seeds. Phytochemistry. 1983;22(11):2359–63.

64. Boye J, Zare F, Pletch A. Pulse proteins: Processing, characterization, functional properties and applications in food and feed. Food Research International. 2010;43(2):414–31.

65. Geerts MEJ, Strijbos M, van der Padt A, van der Goot AJ. Understanding functional properties of mildly refined starch fractions of yellow pea. J Cereal Sci. 2017;75:116–23.

66. Pelgrom PJM, Boom RM, Schutyser MAI. Functional analysis of mildly refined fractions from yellow pea. Food Hydrocoll. 2015;44:12–22.

67. Möller AC, Li J, van der Goot AJ, van der Padt A. A water-only process to fractionate yellow peas into its constituents. IFSET. 2022;75(October 2021).

68. Kornet R, Yang J, Venema P, van der Linden E, Sagis LMC. Optimizing pea protein fractionation to yield protein fractions with a high foaming and emulsifying capacity. Food Hydrocoll. 2022;126(July 2021):107456.

69. Mosenthin R, Messerschmidt U, Sauer N, Carré P, Quinsac A, Schöne F. Effect of the desolventizing/toasting process on chemical composition and protein quality of rapeseed meal. J Anim Sci Biotechnol. 2016;7(1):1–12.

70. Nikiforidis C v., Kiosseoglou V. Aqueous extraction of oil bodies from maize germ (Zea mays) and characterization of the resulting natural oil-in-water emulsion. J Agric Food Chem. 2009;57(12):5591–6.

71. Romero-Guzmán MJ, Jung L, Kyriakopoulou K, Boom RM, Nikiforidis C v. Efficient single-step rapeseed oleosome extraction using twin-screw press. J Food Eng. 2020;276.

72. Abdullah, Weiss J, Zhang H. Recent advances in the composition, extraction and food applications of plant-derived oleosomes. Trends Food Sci Technol. 2020;106(September):322–32.

73. de Chirico S, di Bari V, Foster T, Gray D. Enhancing the recovery of oilseed rape seed oil bodies (oleosomes) using bicarbonate-based soaking and grinding media. Food Chem. 2018;241(September 2017):419–26.

74. Karefyllakis D, Octaviana H, van der Goot AJ, Nikiforidis C v. The emulsifying performance of mildly derived mixtures from sunflower seeds. Food Hydrocoll. 2019;88(June 2018):75–85.

75. Ntone E, Bitter JH, Nikiforidis C v. Not sequentially but simultaneously: Facile extraction of proteins and oleosomes from oilseeds. Food Hydrocoll. 2020;102:105598.

76. Romero-Guzmán MJ, Petris V, de Chirico S, di Bari V, Gray D, Boom RM, et al. The effect of monovalent (Na+, K+) and divalent (Ca2+, Mg2+) cations on rapeseed oleosome (oil body) extraction and stability at pH 7. Food Chem. 2020;306(July 2019):125578.

77. Berghout JAM, Venema P, Boom RM, van der Goot AJ. Comparing functional properties of concentrated protein isolates with freeze-dried protein isolates from lupin seeds. Food Hydrocoll. 2015;51:346–54.

78. Yang Q, Eikelboom E, van der Linden E, de Vries R, Venema P. A mild hybrid liquid separation to obtain functional mungbean protein. LWT. 2022;154(April 2021):112784.

79. Avila Ruiz G, Arts A, Minor M, Schutyser M. A Hybrid Dry and Aqueous Fractionation Method to Obtain Protein-Rich Fractions from Quinoa (Chenopodium quinoa Willd). Food Bioproc Tech. 2016;9(9):1502–10.

80. Diedericks CF, Shek C, Jideani VA, Venema P, van der Linden E. Physicochemical properties and gelling behaviour of Bambara groundnut protein isolates and protein-enriched fractions. Int Food Res J. 2020;138(PB):109773.

81. Johansson M, Johansson D, Ström A, Rydén J, Nilsson K, Karlsson J, et al. Effect of starch and fibre on faba bean protein gel characteristics. Food Hydrocoll. 2022;131(December 2021).

82. Cheung L, Wanasundara J, Nickerson MT. The Effect of pH and NaCl Levels on the Physicochemical and Emulsifying Properties of a Cruciferin Protein Isolate. Food Biophys. 2014;9(2):105–13.

83. Yang J, Berton-Carabin CC, Nikiforidis C v., van der Linden E, Sagis LMC. Competition of rapeseed proteins and oleosomes for the air-water interface and its effect on the foaming properties of protein-oleosome mixtures. Food Hydrocoll. 2021;122:107078.

84. Hinderink EBA, Sagis L, Schroën K, Berton-Carabin CC. Behavior of plant-dairy protein blends at air-water and oil-water interfaces. Colloids Surf B Biointerfaces. 2020;192:111015.

85. Yang J, de Wit A, Diedericks CF, Venema P, van der Linden E, Sagis LMC. Foaming and emulsifying properties of extensively and mildly extracted Bambara groundnut proteins: a comparison of legumin, vicilin and albumin protein. Food Hydrocoll. 2022;123:107190.

86. Dumoulin L, Jacquet N, Malumba P, Richel A, Blecker C. Dry and wet fractionation of plant proteins: How a hybrid process increases yield and impacts nutritional value of faba beans proteins. IFSET. 2021;72(May):102747.

87. Sari YW, Mulder WJ, Sanders JPM, Bruins ME. Towards plant protein refinery: Review on protein extraction using alkali and potential enzymatic assistance. Biotechnol J. 2015;10(8):1138–57.

88. Pernollet JC. Protein bodies of seeds: Ultrastructure, biochemistry, biosynthesis and degradation. Phytochemistry. 1978;17(9):1473–80.

89. Möller AC, van der Padt A, van der Goot AJ. From raw material to mildly refined ingredient – Linking structure to composition to understand fractionation processes. J Food Eng. 2021;291.

90. Funke M, Boom RM, Weiss J. Dry fractionation of lentils by air classification - Composition, interfacial properties and behavior in concentrated O/W emulsions. LWT. 2022;154:112718.

91. Loveday SM. Plant protein ingredients with food functionality potential. Nutr Bull. 2020 Sep 1;45(3):321–7.

92. Gonzalez-Perez S, Vereijken JM, van Koningsveld GA, Gruppen H, Voragen AGJ. Physicochemical Properties of 2S albumins and the corresponding protein isolate from Sunflower (Helianthus annuus). FCT. 2005;70(1):C98–103.

93. Yang J, Kornet R, Diedericks CF, Yang Q, Berton-Carabin CC, Nikiforidis C v., et al. Rethinking plant protein extraction: Albumin—From side stream to an excellent foaming ingredient. Food Struct. 2022;31(January).

94. Kornet R, Penris S, Venema P, van der Goot AJ, Meinders MBJ, van der Linden E. How pea fractions with different protein composition and purity can substitute WPI in heat-set gels. Food Hydrocoll. 2021;120:106891.

95. Wong A, Pitts K, Jayasena V, Johnson S. Isolation and foaming functionality of acid-soluble protein from lupin (Lupinus angustifolius) kernels. J Sci Food Agric. 2013;93(15):3755–62.

96. Ntone E, Wesel T van, Sagis LMC, Meinders M, Bitter JH, Nikiforidis C v. Adsorption of rapeseed proteins at oil / water interfaces . Janus-like napins dominate the interface. J Colloid Interface Sci. 2021;583:459–69.

97. Chua JY, Liu SQ. Soy whey: More than just wastewater from tofu and soy protein isolate industry. Trends Food Sci Technol. 2019;91:24–32.

98. Geerts MEJ, Nikiforidis C v., van der Goot AJ, van der Padt A. Protein nativity explains emulsifying properties of aqueous extracted protein components from yellow pea. Food Struct. 2017;14:104–11.

99. Kornet R, Veenemans J, Venema P, van der Goot AJ, Meinders M, Sagis L, et al. Less is more: Limited fractionation yields stronger gels for pea proteins. Food Hydrocoll. 2021;112:106285.

100. Nikbakht Nasrabadi M, Sedaghat Doost A, Mezzenga R. Modification approaches of plant-based proteins to improve their techno-functionality and use in food products. Food Hydrocoll. 2021;118:106789.

101. Keppler JK, Schwarz K, van der Goot AJ. Covalent modification of food proteins by plant-based ingredients (polyphenols and organosulphur compounds): A commonplace reaction with novel utilization potential. Trends Food Sci Technol. 2020;101:38–49.

102. Shahidi F, Senadheera R. Encyclopedia of Food Chemistry: Protein–Phenol Interactions. In: Encyclopedia of Food Chemistry. Elsevier; 2019. p. 532–8.

103. Tzen JTC, Huang AHC. Surface structure and properties of plant seed oil bodies. Journal of Cell Biology. 1992;117(2):327–35.

104. Ding J, Wen J, Wang J, Tian R, Yu L, Jiang L, et al. The physicochemical properties and gastrointestinal fate of oleosomes from non-heated and heated soymilk. Food Hydrocoll. 2020;100(October 2019):105418.

105. Ding J, Xu Z, Qi B, Liu Z, Yu L, Yan Z, et al. Thermally treated soya bean oleosomes: the changes in their stability and associated proteins. Int J Food Sci Technol. 2020;55(1):229–38.

106. Kapchie VN, Yao L, Hauck CC, Wang T, Murphy PA. Oxidative stability of soybean oil in oleosomes as affected by pH and iron. Food Chem. 2013;141(3):2286–93.

107. Fisk ID, Linforth R, Trophardy G, Gray D. Entrapment of a volatile lipophilic aroma compound (d-limonene) in spray dried water-washed oil bodies naturally derived from sunflower seeds (Helianthus annus). Int Food Res J. 2013;54(1):861–6.

108. Ishii T, Matsumiya K, Nambu Y, Samoto M, Yanagisawa M, Matsumura Y. Interfacial and emulsifying properties of crude and purified soybean oil bodies. Food Struct. 2017;12:64–72.

109. Romero-Guzmán MJ, Köllmann N, Zhang L, Boom RM, Nikiforidis C v. Controlled oleosome extraction to produce a plant-based mayonnaise-like emulsion using solely rapeseed seeds. LWT. 2020;123(August 2019):109120.

110. Bibat MAD, Ang MJ, Eun J bang. Impact of replacing pork backfat with rapeseed oleosomes – Natural pre-emulsified oil – On technological properties of meat model systems. Meat Sci. 2022;186(January):108732.

111. Ntone E, Kornet R, Venema P, Meinders MBJ, van der Linden E, Bitter JH, et al. Napins and cruciferins in rapeseed protein extracts have complementary roles in structuring emulsion-filled gels. Food Hydrocoll. 2022;125(November 2021):107400.

112. Yang J, Faber I, Berton-Carabin CC, Nikiforidis C v., van der Linden E, Sagis LMC. Foams and air-water interfaces stabilised by mildly purified rapeseed proteins after defatting. Food Hydrocoll. 2020;112:106270.

113. Karefyllakis D, Altunkaya S, Berton-Carabin CC, van der Goot AJ, Nikiforidis C v. Physical bonding between sunflower proteins and phenols: Impact on interfacial properties. Food Hydrocoll. 2017;73:326–34.

114. Martins ZE, Pinho O, Ferreira IMPLVO. Food industry by-products used as functional ingredients of bakery products. Trends Food Sci Technol. 2017;67:106–28.

115. Subaşı BG, Vahapoğlu B, Capanoglu E. A review on protein extracts from sunflower cake : techno-functional properties and promising modification methods. Crit Rev Food Sci Nutr. 2021;0(0):1–16.

116. Geerts MEJ, Mienis E, Nikiforidis C v., van der Padt A, van der Goot AJ. Mildly refined fractions of yellow peas show rich behaviour in thickened oil-in-water emulsions. IFSET. 2017;41:251–8.

117. Sahin AW, Hardiman K, Atzler JJ, Vogelsang-O'Dwyer M, Valdeperez D, Münch S, et al. Rejuvenated Brewer's Spent Grain: The impact of two BSG-derived ingredients on techno-functional and nutritional characteristics of fibre-enriched pasta. IFSET. 2021;68(December 2020):102633.

118. Asgar MA, Fazilah A, Huda N, Bhat R, Karim AA. Nonmeat protein alternatives as meat extenders and meat analogs. Compr Rev Food Sci Food Saf. 2010;9(5):513–29.

119. Amin A, Petersen IL, Malmberg C, Orlien V. Perspective on the Effect of Protein Extraction Method on the Antinutritional Factor (ANF) Content in Seeds. ACS Food Science and Technology. 2022;2(4):604–12.

120. Zisopoulos FK, Rossier-Miranda FJ, van der Goot AJ, Boom RM. The use of exergetic indicators in the food industry – A review. Crit Rev Food Sci Nutr. 2017;57(1):197–211.

121. Guinée JB. Handbook on Life Cycle Assessment: operational guide to the ISO standards. Vol. 7. Lindeijer E, editor. Kluwer Academic Publischers; 2002.

122. Grommers HE, van der Krogt DA. Potato Starch: Production, Modifications and Uses. In: Starch. Third Edit. Elsevier Inc.; 2009. p. 511–39.

123. Berghout JAM, Boom RM, van der Goot AJ. The potential of aqueous fractionation of lupin seeds for high-protein foods. Food Chem. 2014;159:64–70.

124. Romero-Guzmán MJ. Designing a sustainable oleosome aqueous extraction: A new way to make emulsion-based foods [Internet]. 2020. Available from: https://research.wur.nl/en/publications/designing-a-sustainable-oleosome-aqueous-extraction-a-new-way-to-

125. Hetherington AC. Life Cycle Assessment of the Production of Edible Oil Emulsions - Comparing a Novel Process Route Using Aqueously Extracted Oil-Bodies Against Existing Technology. 2014;

126. Heusala H, Sinkko T, Sözer N, Hytönen E, Mogensen L, Knudsen MT. Carbon footprint and land use of oat and faba bean protein concentrates using a life cycle assessment approach. J Clean Prod. 2020;242:1–9.

127. Depping V, Grunow M, van Middelaar C, Dumpler J. Integrating environmental impact assessment into new product development and processing-technology selection : Milk concentrates as substitutes for milk powders. J Clean Prod. 2017;149:1–10.

128. Peng Y, Kersten N, Kyriakopoulou K, van der Goot AJ. Functional properties of mildly fractionated soy protein as influenced by the processing pH. J Food Eng. 2020;275:109875.

129. Lie-Piang A, Möller AC, Köllmann N, Garre A, Boom RM, van der Padt A. Functionality-driven food product formulation – An illustration on selecting sustainable ingredients building viscosity. Int Food Res J. 2022;152(December 2021).

130. Batista LF, Marques CS, Pires AC dos S, Minim LA, Soares N de FF, Vidigal MCTR. Artificial neural networks modeling of non-fat yogurt texture properties: effect of process conditions and food composition. Food and Bioproducts Processing. 2021;126:164–74.

131. Passe D, Fouache C, Verrin JM;, Bureau S. PEA PROTEIN COMPOSITION. 2008;

132. Wäsche A, Holley W, Luck T, Nurrenbach T, Borcherding A. Method for Treating and Processing Lupine Seeds Containing Alkaloid, Oil and Protein. 2002;

133. Sridharan S, Meinders MBJ, Bitter JH, Nikiforidis C v. Pea flour as stabilizer of oil-in-water emulsions: Protein purification unnecessary. Food Hydrocoll. 2020;101:105533.

134. Berardy A, Costello C, Seager TP. Life Cycle Assessment of Soy Protein Isolate. Proceedings of the International Symposium on Systainable Systems and Technologies. 2015;v3(August 2017).

135. Vogelsang-o'Dwyer M, Petersen IL, Joehnke MS, Sørensen JC, Bez J, Detzel A, et al. Comparison of Faba Bean Protein Ingredients Environmental Performance. Foods. 2020;9:322.

136. Adebiyi AP, Aluko RE. Functional properties of protein fractions obtained from commercial yellow field pea (Pisum sativum L.) seed protein isolate. Food Chem. 2011;128(4):902–8.

137. Holt NW, Sosulski FW. Amino Acid Composition and Protein Quality of Field Peas. Canadian Journal of Plant Science. 1979;59(3):653–60.

138. van Paassen M, Braconi N, Kulling L, Durlinger B, Gual P. Agri-footprint 5.0 - Part 2: Description of data. Gouda; 2019.

139. Snowden J, Sipsas S, Charmian J. Method to produce lupin protein-based dairy substitutes. 2007;1(19).

140. Huijbregts M, Steinmann ZJN, Elshout PMFM, Stam G, Verones F, Vieira MDM, et al. ReCiPe 2016 A harmonized life cycle impact assessment method at midpoint and endpoint level Report I: Characterization. National Institute for Public Health and the Environment [Internet]. 2016;194. Available from: https://www.rivm.nl/bibliotheek/rapporten/2016-0104.pdf

141. Fantke P, Aurisano N, Bare J, Backhaus T, Bulle C, Chapman PM, et al. Toward harmonizing ecotoxicity characterization in life cycle impact assessment. Environ Toxicol Chem. 2018;37(12):2955–71.

142. Berghout JAM, Boom RM, van der Goot AJ. The potential of aqueous fractionation of lupin seeds for high-protein foods. Food Chem [Internet]. 2014;159:64–70. Available from: http://dx.doi.org/10.1016/j.foodchem.2014.02.166

143. European Commission. Guidance 19: Vegetable oil and animal fat extraction and vegetable oil refining activities. Guidance on VOC Substitution and reduction for activities covered by the VOC solvents emissions directive (Directive 1999/13/EC). 2008;

144. Thrane M, Paulsen P v., Orcutt MW, Krieger TM. Soy Protein: Impacts, Production, and Applications. Sustainable Protein Sources. 2017;23–45.

145. Servais C, Jones R, Roberts I. The influence of particle size distribution on the processing of food. J Food Eng. 2002;51(3):201–8.

146. Dekkers BL, de Kort DW, Grabowska KJ, Tian B, van As H, van der Goot AJ. A combined rheology and time domain NMR approach for determining water distributions in protein blends. Food Hydrocoll. 2016;60:525–32.

147. González-Tejedor GA, Martínez-Hernández GB, Garre A, Egea JA, Fernández PS, Artés-Hernández F. Quality Changes and Shelf-Life Prediction of a Fresh Fruit and Vegetable Purple Smoothie. Food Bioproc Tech. 2017;10(10):1892–904.

148. Manzocco L, Calligaris S, Camerin M, Pizzale L, Nicoli MC. Prediction of firmness and physical stability of low-fat chocolate spreads. J Food Eng. 2014;126:120–5.

149. Monnet AF, Michon C, Jeuffroy MH, Blumenthal D. Taking into Account Upstream Variability of Flours with Processing Variables in Legume-Enriched Soft Cakes: Conception of a Multiobjective Model for the Monitoring of Physical Properties. Food Bioproc Tech. 2019;12(4):625–35.

150. la Fé Perdomo I, Quiza R, Haeseldonckx D, Rivas M. Sustainability-Focused Multi-objective Optimization of a Turning process. Int J Precis Eng Manuf - Green Technology. 2019;7(5):1009–18.

151. Jonkman J, Kanellopoulos A, Bloemhof JM. Designing an eco-efficient biomass-based supply chain using a multi-actor optimisation model. J Clean Prod. 2019;210:1065–75.

152. Mccleary B v, Cox J. Evolution of a Definition for Dietary Fiber and Methodology to Service this Definition. Luminacoids Research. 2017;21(2):9–21.

153. Balet S, Guelpa A, Fox G, Manley M. Rapid Visco Analyser (RVA) as a Tool for Measuring Starch-Related Physiochemical Properties in Cereals: a Review. Food Anal Methods. 2019;12(10):2344–60.

154. Box GEP, Hunter JS, Hunter GW. Statistics for Experimenters: Design, Innovation, and Discovery. Second Edi. Hoboken, N.J.: John Wiley & Sons; 2005.

155. Bravo-Núñez Á, Garzón R, Rosell CM, Gómez M. Evaluation of starch-protein interactions as a function of pH. Foods. 2019;8(5):2–11.

156. Căpriţă A, Căpriţă R, Simulescu VO, Drehe RM. The Effect of Temperature on Soluble Dietary Fiber Fraction in Cereals The Effect of Temperature on Soluble Dietary Fiber Fraction in Cereals. J Agroaliment Processes Technol. 2011;17(3):214–7.

157. Korson L, Drost-Hansen W, Millero FJ. Viscosity of water at various temperatures. Journal of Physical Chemistry. 1969;73(1):34–9.

158. Lam ACY, Can Karaca A, Tyler RT, Nickerson MT. Pea protein isolates: Structure, extraction, and functionality. Food Rev Int. 2018;34(2):126–47.

159. Ma Y, Schutyser MAI, Boom RM, Zhang L. Predicting the extrudability of complex food materials during 3D printing based on image analysis and gray-box data-driven modelling. IFSET. 2021;73(July):102764.

160. Geerts MEJ, Dekkers BL, van der Padt A, van der Goot AJ. Aqueous fractionation processes of soy protein for fibrous structure formation. IFSET. 2018;45(December 2017):313–9.

161. Urbonaite V, de Jongh HHJ, van der Linden E, Pouvreau L. Origin of water loss from soy protein gels. J Agric Food Chem. 2014;62(30):7550–8.

162. Kuhn M. caret: Classification and Regression Training [Internet]. 2022. Available from: https://github.com/topepo/caret/

163. Claesen M, De Moor B. Hyperparameter Search in Machine Learning. 2015;10–4.

164. Draper NR, Smith H. Applied Regression Analysis. Third edit. John Wiley & Sons; 1998.

165. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.

166. Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;

167. Breiman L. Random Forest. Mach Learn. 2001;45:5–32.

168. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Vol. 2. New York: Springer; 2009.

169. Kvålseth TO. Cautionary note about r2. American Statistician. 1985;39(4):279–85.

170. Sirtori E, Resta D, Brambilla F, Zacherl C, Arnoldi A. The effects of various processing conditions on a protein isolate from Lupinus angustifolius. Food Chem. 2010;120(2):496–504.

171. Möller AC, van der Padt A, van der Goot AJ. Influence of the fractionation method on the protein composition and functional properties. IFSET. 2022;81(May).

172. Brownlee J. Better deep learning: train faster, reduce overfitting, and make better predictions. Machine Learning Mastery; 2018.

173. Ghojogh B, Crowley M. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. 2019;1–23. Available from: http://arxiv.org/abs/1905.12787

174. Möller AC, Lie-Piang A, Bian T, van der Padt A, van der Goot AJ. Structure analysis of multi-component biopolymer blends – The effect of water distribution on the viscoelastic properties (Submitted). 2022;

175. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. 2021.

176. Milborrow S. earth: Multivariate Adaptive Regression Splines. 2021.

177. Breiman L, Cutler A, Liaw A, Wiener M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. 2022.

178. Fritsch S, Guenther F, Wright MN. neuralnet: Training of Neural Networks. 2019.

179. Lie-Piang A, Garre A, Nissink T, van Beek N, van der Padt A, Boom R. Machine learning to quantify techno-functional properties - A case study for gel stiffness with pea ingredients. IFSET. 2023 Jan;83:103242.

180. Elleuch M, Bedigian D, Roiseux O, Besbes S, Blecker C, Attia H. Dietary fibre and fibre-rich by-products of food processing: Characterisation, technological functionality and commercial applications: A review. Vol. 124, Food Chem. 2011. p. 411–21.

181. Hinderink EBA, Schröder A, Sagis L, Schroën K, Berton-Carabin CC. Physical and oxidative stability of food emulsions prepared with pea protein fractions. Lwt. 2021;146(March).

182. Zhou X, Sala G, Sagis LMC. Bulk and interfacial properties of milk fat emulsions stabilized by whey protein isolate and whey protein aggregates. Food Hydrocoll. 2020;109(June):106100.

183. Lie-Piang A. Functionality-oriented selection of ingredients [Internet]. 2022. Available from: https://git.wur.nl/anouk.lie-piang/toward-ml-for-food-formulation

184. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees [Internet]. 2019. Available from: https://cran.r-project.org/package=rpart

185. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica . 2012;22(3):276–82.

186. Kennard RW, Stone LA. Computer Aided Design of Experiments. Vol. 11. 1969.

187. Hageman JA, Engel B, de Vos RCH, Mumm R, Hall RD, Jwanro H, et al. Robust and Confident Predictor Selection in Metabolomics. In: Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry. Springer International Publishing; 2017. p. 239–57.

188. Singh N. Functional and physicochemical properties of pulse starch. In: Pulse Foods: Processing, Quality and Nutraceutical Applications. Elsevier; 2020. p. 87–112.

189. McClements DJ. Food Emulsions Principles, Practices, and Techniques. Second Edi. CRC Press, Boca Raton, FL.; 2005.

190. Townsend AA, Nakai S. Relationships Between Hydrophobicity and Foaming Characteristics of Food Proteins. J Food Sci. 1983;48(2):588–94.

191. Osen R, Toelstede S, Wild F, Eisner P, Schweiggert-Weisz U. High moisture extrusion cooking of pea protein isolates: Raw material characteristics, extruder responses, and texture properties. J Food Eng. 2014 Apr;127:67–74.

192. Hawkins DM. The Problem of Overfitting. Vol. 44, J.Chem.Inf.Comput.Sci. 2004. p. 1–12.

193. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Join Conference on Artifical Intelligence . 1995.

194. Becker D, Schmitt C, Bovetto L, Rauh C, McHardy C, Hartmann C. Optimization of complex food formulations using robotics and active learning. IFSET. 2023 Jan;83:103232.

Summary

Food products like vegan yoghurts or ready-to-eat soups require texturizing ingredients, such that the yoghurt becomes stiff and the soup becomes thick. These texturizing ingredients are commonly produced from crops such as soybeans or peas. Certain components like protein or starch that are known to provide this texture to the products are separated from these crops. Conventionally, crops are fractionated into ingredients of high purity, also called isolates. In this way, they always have the same composition and therefore the same functionality in each batch. However, this process requires many resources such as water to solubilize the crops but also energy to subsequently dry the ingredients again. In addition, harsh processing conditions like the addition of chemicals or the application of heat may change the properties of the components in the ingredients.

Alternatively, milder fractionation methods do not refine the crops into pure ingredients. For example, peas can be milled into flour and then separated using airflow in an air classifier, which separates the particles based on size and density. This does result in highly functional ingredients, but with lower purity. An overview of milder fractionation methods to produce ingredients rich in among others protein and oil is provided in **chapter 2 – state of the art.** The production of milder refined ingredients requires substantially fewer resources, which was quantified for a selection of these methods in chapter **3 – sustainability.** Even though some milder refined ingredients are less pure, it was found that these have a lower environmental impact when expressed per kg of protein.

However, as these milder refined ingredients are less pure, the functionality is understood less well compared to pure ingredients. The complex composition therefore limits their application in food products. To facilitate their use in food products, an approach inspired by the method of process system engineering can be applied. This is for example used to blend chemicals into a final product: ingredients are blended together to match the composition of food products. Although with this approach milder refined ingredients could be matched to final food products, it resulted in a substantial amount of ingredients that could not be matched due to the normally unwanted components like fibre. Therefore, more raw materials were needed and the overall sustainability was suboptimal.

This thesis assesses whether milder refined ingredients can be selected based on their techno-functional properties instead of their compositions, which would result in more sustainable food products. An example is to select an ingredient on its capacity to increase the gel stiffness or to increase the thickness of the soup instead of the amount of protein and starch present. This was done using a case study in which conventional highly refined and milder refined air-classified yellow pea and lupine seed ingredients were matched to a product portfolio. As a proof of concept, four main techno-functional properties were quantified using machine learning, namely the thickening behaviour (pasting and gelling), foaming capacity, and emulsion stabilization.

The concept of a functionality-driven ingredient selection was first illustrated in **chapter 4 – concept**. Here, the thickening capacity of yellow pea ingredients was quantified using multiple linear regression with the composition (protein, starch, fibre, and the rest) as input variables. With this relationship, the formulation space to reach a certain viscosity could be revealed. In this space, ingredients could now be selected and optimised for their composition and environmental impact. To extend this concept to non-linear techno-functional properties, other machine learning models would have to be employed that can model this behaviour. Therefore, in **chapter 5 – framework,** we presented a framework in which machine learning models can be selected to quantify gel stiffness while applying expert knowledge. This means that the selected model had to give plausible physical predictions without artefacts that could complicate the formulation of ingredients.

Most ingredient formulations consist of ingredients originating from multiple crops. Therefore, single machine learning models were created that could predict ingredients from both yellow pea and lupine, for all selected techno-functional properties in **chapter 6 – mixed crops**. The models based on individual crops were compared to the models based on all data, which yielded in most cases a similar prediction accuracy, with sometimes a trade-off of higher errors.

Finally, in **chapter 7 – General Discussion**, we discuss the main findings of this thesis and compile them into the final application: the selection of yellow pea and lupine ingredients based on techno-functional properties for a product portfolio of

14 products. The global warming potential of the portfolio with only conventional ingredients could be reduced by 50% by also matching milder refined ingredients based on composition. However, these formulations should be taken with caution as they do not always result in the desired functional properties. A 70% reduction in global warming can be achieved by matching these ingredients based on techno-functional properties. The further reduction is attributed to the larger formulation space that is revealed by the functional properties and to the fact that milder refined ingredients deliver in most cases more functionality per kg CO_2 – equivalents (i.e., global warming potential).

This thesis demonstrates that machine learning has a high potential to describe techno-functional properties, to incorporate more complex but more sustainable ingredients into food product formulations. Future research should now focus on improving and extending the created models to other functional properties and processing conditions while the data collection could be made more efficient.

Appendices

About the author

Anouk Lie-Piang was born in Rotterdam, the Netherlands, on July 5, 1994. In high school, at the Montessori Lyceum Rotterdam, she was quickly drawn to the sciences, inspiring her decision to study liberal arts and sciences at Amsterdam University College. First introduced to the field of food science in the course Gastronomy, she went on to spend a semester at McGill University in Montreal, where she developed a keen interest in food technology. To gain more practical experience, Anouk interned at a Gouda cheese factory and learned about the several processing steps involved in cheesemaking. In 2016, she began the European Master of Food Science, a master program of leading universities in four European countries. In her first year, she completed courses and studied at Wageningen University & Research, University College Cork, AgroParisTech, and Lund University. In her second year, she did a combined thesis internship project at Tetra Pak in Lund, focusing on the behaviour of milk proteins in skim milk powder during reconstitution. After completing this project, she stayed on as a consultant at Tetra Pak, further developing her research and project management skills. In 2019, Anouk continued working as a PhD candidate at the Food Process Engineering group at Wageningen University & Research on the project titled 'Functionality-driven food formulation - Reducing the environmental impact using machine learning'.

Written by my dear friend Anne van den Bergh.

Acknowledgements

While reflecting on my PhD journey, I am feeling very grateful to everyone who has contributed to this thesis. Either by sharing knowledge or challenging me in discussions. But also by making this whole journey more fun with small talks at the coffee machine and celebrations of all small and big achievements.

I would like to start by thanking my promotors Remko and Albert, both of you are a source of inspiration. It is your creativity and forward thinking that have made this project possible, which was a bit of an experiment from the beginning. Thanks for giving me the freedom to create this project and letting me develop the skills that I would like to have, whilst always encouraging me and subtly steering the project in the right direction.

Thanks to all the technicians who have supported the students I supervised and me with the lab and computer work, always with a smile and patience. Martin, thanks for the financial support and your patience when I was late again with filling in the hours. But also for the small spot in your lab to use the foaming set-up. Jos, thanks for all your help, mainly when I was figuring out the mill again. Jarno, thanks for helping me out whilst developing the tubes for creating gels. Wouter, thanks for measuring yet another Dumas sample for me and giving the final months of my PhD a boost with the extra computing power. Maurice, thanks for your help with the microscopes and for always participating in the Stukjes I made. Lyneth, we never got the chance to work together, but thank you for the fun talks anyhow. I would also like to thank Harry for helping me out with some equipment in FPH.

Besides lab and computer work, I am also grateful for the drivers behind the well-oiled secretary of FPE. Marjan, Ilona, Chantal, and Evelyn for all your help with booking, organising, and setting meetings.

I strongly believe that projects benefit from fruitful (interdisciplinary) collaborations, which I have had the luck of experiencing during my PhD. I have received a tremendous amount of help from Alberto Garre from FHM. With your help, I could steer my PhD project in the direction of data science, which ultimately

shaped the complete project. Thanks for making time for me, becoming co-author of two papers, always replying when I got stuck with my R code, and for becoming a friend of mine. Jos Hageman from Biometris has also contributed greatly to a part of the statistical backbone of my work. Thanks for being so open to collaborating only after we met when I took your Chemometrics course. Nicolò from Blonk Consultants, thanks for introducing me to the world of life cycle assessments and Simapro. Anna, Andrea, and Nynke, thanks for showing me it is possible to have inspiring collaborations between colleagues and friends, which ultimately resulted in three papers together. Maarten, Costas, and Jack, thanks for your valuable input in our review paper.

The collaborations with BSc and MSc students have been an essential part of this work as well. Anaïs, Nienke, Thomas, and Eline, thanks for participating so enthusiastically at the beginning of my PhD and helping me find the right direction in my thesis, which was at that time still very vague. Niels, Bente, Ting, Zeger, Vincent, Iris, Kai, and Suzan thank you for being excited about this topic and for your tremendous (lab work) contribution. To all, thank you for your motivation and curiosity, it was a joy seeing you develop as researchers during your theses.

Next to lab work and meetings, I have always had a warm office to come back to. At first, this started in the extra FPE office in the FPH group with Zaida, Tatsuya, Eric, and Sten. Thanks for creating a fun office even though we were a bit far from home. But also thanks to Raisa and Elke, who welcomed me warmly to the (many) infamous FPH borrels with a lot of beer pong and French cheese. After, I got a desk in the most cheerful and decorated office in FPE with Andrea, Nynke, and Lina. Thanks for the good conversations and laughs when we were yet again standing on chairs and tables to hang up lights, plants or a Christmas star. Later Iris, Kübra, Lingfeng, Fathinah, and Ana also joined us, who all didn't seem to mind being in an office in which is it all year long Día de los Muertos and Christmas at the same time. Thanks for creating a cosy atmosphere in the office and for always being up for a fun chat and office dinners.

Of course, the fun continued outside of my office as well. Thanks to my PhD-partner in crime Rebecca, who has stood by my side from the beginning until the end. I am

so grateful that we have met and experienced this journey together, sipping coffee in *our* green chairs contemplating about life or why there was no milk anymore for our favorite milky shrimp dish. Thanks to all my FPE colleagues, for the small talks in the hallways, lab-uitjes, PhD trip to Singapore, coffee breaks, and fun borrels. Anna, Raisa, Nynke, Jan, Emma, Floor, and Steven thanks for all the fun at the borrels in and outside of FPE in my first year. At this time, I was almost feeling like a first year Bachelor student again. Yizhou, thanks for the good conversations, drinks, laughs, yet also fruitful discussions where you always challenge me. Yifeng, Hilda, and Andrea, for being great travel buddies in Thailand with our great guides Polly and Nattawan showing us a part of your home country. Anna, Eric, Hilda, and later also Zulhaj for making the ISPT project meeting more fun. Pina, for our good talks and your support during the first months of my PhD. Jan-Eise for the fun chats and your great baking skills. Thanks Sten for organising the Veluweloop and to everyone who participated, it was a great day. Andrea and Anna, thanks for creating a safe home for me in Wageningen filled with laughter, tasty food, drinks, and discussions. And of course thanks to all my other colleagues for the nice chats, food, and fun we had: Wanqing, Isabel, Juliana, Ruoxuan, Yu, Lu, Bijoy, Regina, Ivanna, Meinou, Julia, Claudine, Patrick, Regina, Bijoy, Luc, Yafei, Elvira, Konstantina, Joanne, Murat, Anneloes, Koen, Julia, Miek, Silvia, Ruihao, Quinten, Katharina, Mingzhao, Loes, Sirinan, Dan, Jolien, Kieke, Sybren, Aadi, Laurens, and Asif.

Tot slot wil ik ook nog mijn familie en vrienden buiten Wageningen bedanken. Lieve vrienden, door jullie enthousiasme en ondersteuning zijn de afgelopen vier jaar zoveel leuker geworden. Door de vele etentjes, borrels en tripjes kon ik altijd ontspannen en op andere dingen focussen. Lieve Hanna, Charlotte en mama, bedankt voor jullie onuitputtelijke ondersteuning, motivering en natuurlijk vooral liefde. Jullie geloof in mij geeft mij altijd net het duwtje in mijn rug dat ik nodig heb. Tyler, Zee, and Dirk, thanks for your continuous support throughout these four years and sharing your knowledge and beautiful wines with me. Lieve Dimi, dank voor dag en nacht voor mij klaar staan, motiveren en mij een plekje laten claimen in jouw kantoortje. Jouw natuurlijke rust en relativeringsvermogen zijn een essentiële bijdrage geworden aan mijn werk. Ik ben ongelofelijk dankbaar voor alle liefde die ik van jullie allemaal elke dag weer mag ervaren.

Publications

This thesis

Lie-Piang A, Braconi N, Boom RM, van der Padt A. Less refined ingredients have lower environmental impact–A life cycle assessment of protein-rich ingredients from oil-and starch-bearing crops. J. Clean. Prod. 2021 Apr 10;292:126046.

Lie-Piang A, Möller AC, Köllmann N, Garre A, Boom R, van der Padt A. Functionality-driven food product formulation–An illustration on selecting sustainable ingredients building viscosity. Food Res. Int. 2022 Feb 1;152:110889.

Lie-Piang A, Garre A, Nissink T, van Beek N, van der Padt A, Boom R. Machine learning to quantify techno-functional properties-A case study for gel stiffness with pea ingredients. IFSET. 2023 Jan 1;83:103242.

Lie-Piang A, Yang J, Schutyser MA, Nikiforidis CV, Boom RM. Mild Fractionation for More Sustainable Food Ingredients. Annu. Rev. Food Sci. Technol. 2023 Mar 27;14:473-93.

Lie-Piang A, Hageman J, Vreenegoor I, van der Kolk K, de Leeuw S, van der Padt A, Boom RM. Quantifying techno-functional properties of ingredients from multiple crops using machine learning. *Submitted*.

Lie-Piang A, Boom RM, van der Padt A. Ingredient formulations driven by techno-functional properties. *Submitted*.

Other scientific publications

Lie-Piang A, Leeman M, Castro A, Börjesson E, Nilsson L. Investigating the effect of powder manufacturing and reconstitution on casein micelles using asymmetric flow field-flow fractionation (AF4) and transmission electron microscopy. Food Res. Int. 2021 Jan 1;139:109939.

Lie-Piang A, Leeman M, Castro A, Börjesson E, Nilsson L. Revisiting the dynamics of proteins during milk powder hydration using asymmetric flow field-flow fractionation (AF4). CRFS. 2021 Jan 1;4:83-92.

Möller AC, Lie-Piang A, Bian T, van der Padt A, van der Goot AJ. Structure analysis of multi-component pastes–The effect of water distribution on the rheological properties. Food Res. Int. 2022 Dec 1;162:112069.

Draijer N*, Rivera del Rio A*, Lie-Piang A, Janssen AEM, Boom RM. Nutritional value in sustainability assessment of protein-rich ingredients and foods: a 'farm-to-feaces' approach. *Submitted*.

* These authors share first authorship

Overview of completed training activities

Discipline specific activities

Rheology School Leuven	KU Leuven	2019
EFFoST conference (*volunteered*)	Elsevier	2019
Machine Learning by Stanford University	Coursera	2021
EFFost conference[a]	EFFoST	2020
LCA Food Conference[b]	LCA Food	2021
Chemometrics	VLAG	2021
EFFoST conference[b] (Student of the year award Popular Vote)	EFFoST	2021
Aachen Process mining summer school	Aachen University	2022
Food Structure and Functionality Symposium[a]	Elsevier	2022

General courses

WGS PhD Workshop Carousel	WGS	2019
Critical thinking and argumentation	WGS	2019
Brain training	WGS	2019
VLAG PhD Week	VLAG	2019
Introduction to (La)TeX	PE&RC	2019
Introduction to R	VLAG	2020
Applied Statistics	VLAG	2020
Philosophy and Ethics of Food Science and Technology	VLAG	2021
Rmarkdown	WGS	2021
Career perpectives	WGS	2022

Other activities

Preparation of research proposal	Chairgroup	2019
PhD study tour to Singapore[a,b]	Chairgroup	2022
PhD Council	VLAG	2019-2022
Journal club	Chairgroup	2019-2021

[a]oral presentation [b]poster presentation

This thesis was part of the 'Novel Process Routes' project and took place within the framework of the Institute of Sustainable Process Technology (ISPT) and is co-funded by TKI-Energy with the supplementary grant 'TKI- Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy. The authors thanks all members of the ISPT Consortium for their support: Cosun, DSM, FrieslandCampina, Pentair, NIZO, RHDHV, TU Delft, and Wageningen University & Research.

More information can be found at *https://ispt.eu/projects/novel-process-routes/*.

Book and cover design inspired by Anna Cäcilie Möller, Dimitri Roozendaal, and Dirk Jan Postel.

Printed by Proefschriftmaken.nl