



HAL
open science

Crop diversity used in branded products with focus on legume species worldwide

Tristan Salord, Marie-Benoît Magrini, Valérie Lullien-Pellerin, Guillaume Cabanac, Marie-Joseph Amiot, Cecile Barron, Adeline Boire, Valérie Micard, Magalie Weber

► To cite this version:

Tristan Salord, Marie-Benoît Magrini, Valérie Lullien-Pellerin, Guillaume Cabanac, Marie-Joseph Amiot, et al.. Crop diversity used in branded products with focus on legume species worldwide. *npj Science of Food*, 2024, 8 (1), pp.68. 10.1038/s41538-024-00305-7 . hal-04732768

HAL Id: hal-04732768

<https://hal.inrae.fr/hal-04732768v1>

Submitted on 11 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

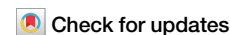
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

<https://doi.org/10.1038/s41538-024-00305-7>

Crop diversity used in branded products with focus on legume species worldwide



Tristan Salord¹✉, Marie-Benoît Magrini¹✉, Valérie Lullien-Pellerin², Guillaume Cabanac^{3,4}, Marie-Joseph Amiot⁵, Cécile Barron², Adeline Boire⁵, Valérie Micard² & Magalie Weber⁶

Food diversity is a challenging issue for sustainable agrifood systems. Diets are increasingly dependent on branded packaged foods. Therefore, the crop diversity offered in the food market through these products is of particular importance. We scrutinize this diversity for some crops under great societal challenge: pulses. Based on the product launches referenced in the Mintel database over the last decade, we compare the food products containing pulse crops with those containing another legume—soy. From the 350,000 products analyzed, our results show that soy is mainly used but reveal some progress in the use of pulse species, particularly in Europe. The position of the examined species in the list of ingredients and in the product description allows us to assess its importance. The text-mining methods used usefully enable the monitoring of crop usage in the food market. We discuss several perspectives, notably how to deepen these results regarding consumer choices.

The environmental impacts of conventional agrifood systems and markets that shape modern diets worldwide are widely acknowledged^{1–3}, particularly in terms of biodiversity loss⁴. At a global scale, crop systems seem to rely on a poor diversity of species⁵. According to the most recent Food and Agriculture Organization of the United Nations (FAO) assessment⁶, out of 6000 different plant species cultivated in 2014, only 9 accounted for 66% of total crop production, and only 3 of them—wheat, rice, and maize—represented more than 50% of plant-based human food. This low level of species diversity of cropping systems worldwide makes crops much more sensitive to pest infestation^{7,8}, less resilient to climate change^{9–12}, and more dependent on synthetic inputs such as fertilizers and pesticides^{13,14}. This situation may also have detrimental consequences on the human diet^{3,15}. Agroecological transition toward greater species diversity and food transition toward more plant-based diets are both required^{12,16,17}, but the related drivers are obviously complex and multidimensional and concern technological, social, institutional, educational, political, and economic levers simultaneously^{18,19}.

Concerning packaged food markets, additional data on food product composition are required to uncover the relationship between species diversity and agrifood systems^{2,16,20}. However, this situation is hampered, on the one hand, by a lack of information in the scientific literature studying the types of ingredients used in food products²¹ and, on the other hand, by a lack of consensus on how to categorize ingredients, notably according to their levels of processing²², beyond the first step of identifying which species are used by the food industry. Our research addresses these challenging

questions by using text-mining methods applied to the ingredient lists of packaged branded foods, regardless of the degree of processing.

Our goal is to question the diversity of species from a market point of view: what crop diversity is used by the food industry throughout the world? To address this question, we focus on food product launches that reflect the innovation dynamics of branded food markets. We assume that the food ingredients in these new products reflect food industry know-how on species usage. According to innovation theories²³ and transition studies^{24,25}, a transition toward more diverse cultivated crops can result from firms investing in new species for their products. In short, we consider the new packaged/processed food offered to be a proxy of the capacity of the food industry to support species diversity.

Given that the mention of the geographical origin of the ingredients used in food products is not mandatory and that international trade exchanges regarding agricultural and food products are becoming increasingly important, we first address this question at the global scale. A deeper analysis to assess this diversity at the regional or country scale by comparing the level of crop diversity observed in food products with that of the main cultivation systems is beyond the scope of this study. The challenge here is to analyze, first, the variety of species through the ingredients used. This diversity of species is one major component of biodiversity issues.

This level of diversity is assessed at the ingredient list level from the following three complementary perspectives: (a) the analysis of the variety of species used, (b) the degree of concentration in the market on certain

¹AGIR, INRAE, Université de Toulouse, Castanet-Tolosan, France. ²IATE, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France. ³IRIT, CNRS, Université de Toulouse, Toulouse, France. ⁴Institut Universitaire de France (IUF), Paris, France. ⁵MOISA, INRAE, Université de Montpellier, Institut Agro, Montpellier, France. ⁶UR BIA, INRAE, F-44316 Nantes, France. ✉e-mail: tristan.salord@inrae.fr; marie-benoit.magrini@inrae.fr

Table 1 | First 10 species frequencies in branded food product launches

Species	Frequ.	Frequ. (%)	Frequ. among top 5 ing.	Frequ. among top 5 ing. (%)	Frequ. in remaining ing. list	Frequ. in remaining ing. list (%)	Frequ. in product description	Frequ. in product description (%)
<i>Glycine max</i>	277,656	71.586	93,359	33.62	184,297	66.38	11,369	4.09
<i>Pisum sativum</i>	36,144	9.319	16,291	45.07	19,853	54.93	10,855	30.03
<i>Phaseolus vulgaris</i>	25,358	6.538	14,668	57.84	10,690	42.16	14,070	55.49
<i>Cicer arietinum</i>	16,160	4.166	11,560	71.53	4600	28.47	7945	49.16
<i>Lens culinaris</i>	9706	2.502	7083	72.98	2623	27.02	6826	70.33
<i>Ceratonia siliqua</i>	7212	1.859	89	1.23	7123	98.77	7	0.1
<i>Phaseolus coccineus</i>	4601	1.186	2398	52.12	2203	47.88	3786	82.29
<i>Vigna radiata</i>	3899	1.005	2005	51.42	1894	48.58	1299	33.32
<i>Lupinus angustifolius</i>	1715	0.442	669	39.01	1046	60.99	455	26.53
<i>Vigna angularis</i>	1284	0.331	698	54.36	586	45.64	381	29.67

The table shows the frequencies of appearance (“Frequ.”) for the ten species most present in the corpus, at different levels: at the corpus level, among the top 5 ingredients, among the remaining items in the ingredient lists, and in the product descriptions. For each of these absolute values, their percentage share is also indicated.

dominant species, and (c) a preliminary approach to species’ contribution to the overall product formulation. These three approaches to determining species diversity in product launches are based on three assumptions. First, the term “diversity” suggests a wide variety of species. If markets feature only a few species, then it would be speculative to speak of diversity. Second, the predominance of a particular species can be a disabling factor for increased levels of cultivated diversity. To assess these effects, we need to look at how food processing markets concentrate on certain species. Third, we hypothesize that the ways in which a species is used by the food industry can either promote or limit species diversity in the market.

Therefore, this paper contributes to the need for better scientific knowledge on the composition of those food products^{26,27} thanks to the development of text-mining methods.

Given the important methodological issues associated with such assessments, we first focus on one group of species under greater societal debate—pulses. Pulses are increasingly promoted as a main lever for both agroecological and food transition^{28,29} from the “one health” perspective. Pulses have benefited from growing public support since the 2016 International Year of the United Nations, which has created great momentum for these plant species³⁰. While pulses are increasingly praised for their contribution to healthy and sustainable agrifood systems, they face a lock-in situation compared to the considerable degree of development of major crops such as wheat and soy^{31–33}. In particular, debates exist on how pulses can develop compared with soy, which is not only already a dominant commodity for feed but also is becoming a dominant commodity for food. Indeed, bibliometric works have shown that more than half of the scientific works on legumes in the food sciences field are on soy^{34,35}. Considering that the development of pulses is a major challenge for food^{28,33}, we choose to compare the diversity of pulse species with that of soy in branded package foods at a global scale.

The remainder of this paper is organized as follows. Based on several statistical figures and tables, the Results section discusses the main insights. The Discussion section enhances the reflection by proposing future works for analyzing the type of processing of those ingredients and hence providing a comprehensive picture of the food technological development of pulses in food markets. We also question how to better organize data and information on food products to further assess the diversity of food species at the country scale and to link our analysis of the food industry supply with the analysis of consumer purchases. The last section (Methods) presents the data retrieved from the Mintel GNPD database and how we parsed the ingredients that were analyzed with food experts to identify the species.

Results

Species diversity in an unbalanced market

The tagging of legume-based ingredients from the dataset led to the identification of 32 different species, including soy (all presented in Table S1 in the Supplementary Materials section and in Table 1 for the first 10 species).

The number of species identified is a rather unexpected observation and, at first glance, can be considered an encouraging result in terms of species diversity. However, we observe a highly asymmetrical distribution of the identified species. Products for which only soy is identified within the ingredient lists account for 73% ($n = 249,425$) of all products ($n = 343,309$). Conversely, products for which only one or more NSL (non-soy legume) ingredients are identified represent 19% ($n = 65,653$) of all products. Eight percent ($n = 28,231$) of the remaining products contain both soy and NSL ingredients. Therefore, nearly 4 times more product launches contain soy ingredients than contain NSL ingredients.

The analysis of the frequencies of these species (Table 1 and Table S1) according to their position confirms this highly unbalanced market in favor of a few species—primarily soy and pea. More precisely, a quartet of species, namely, *Pisum sativum* L., *Phaseolus vulgaris* L., *Cicer arietinum* L., and *Lens culinaris* Medik L. account for almost 79% of NSL ingredients, while approximately twenty NSL species have a frequency of appearance inferior to 1% among products containing NSL species. However, products with NSL ingredients are more frequently associated with the mention of NSL species in the product description (on the packaging) compared with products with soy ingredients, for which this frequency is only 4%. For instance, 70% of products contain *Lens culinaris* Medik. present a mention of the lens species within the product description.

This imbalance between soy and NSL ingredients can be linked to economic factors, such as the availability of each species for food companies. According to the FAOStat, soy remains the most cultivated legume in the world, with an annual production of more than 300 million tons over the last decade. This production level is 3 times greater than the total production of the above-mentioned most frequent NSL species. These observations first confirm the existence of a strong “technological lock-in” around one legume species, soy, which is also widely used for food worldwide, and the dominance of this species can be analyzed as a structural trend in agrifood markets^{29,32}.

This dominance of soy is observed across all market segments (Fig. 1), except the “spreads” and “fruits and vegetables” segments, where the balance between products containing soy ingredients and products containing NSL ingredients is almost negligible or even reversed, respectively. However, the growth rates of products containing soy and NSL ingredients lead us to

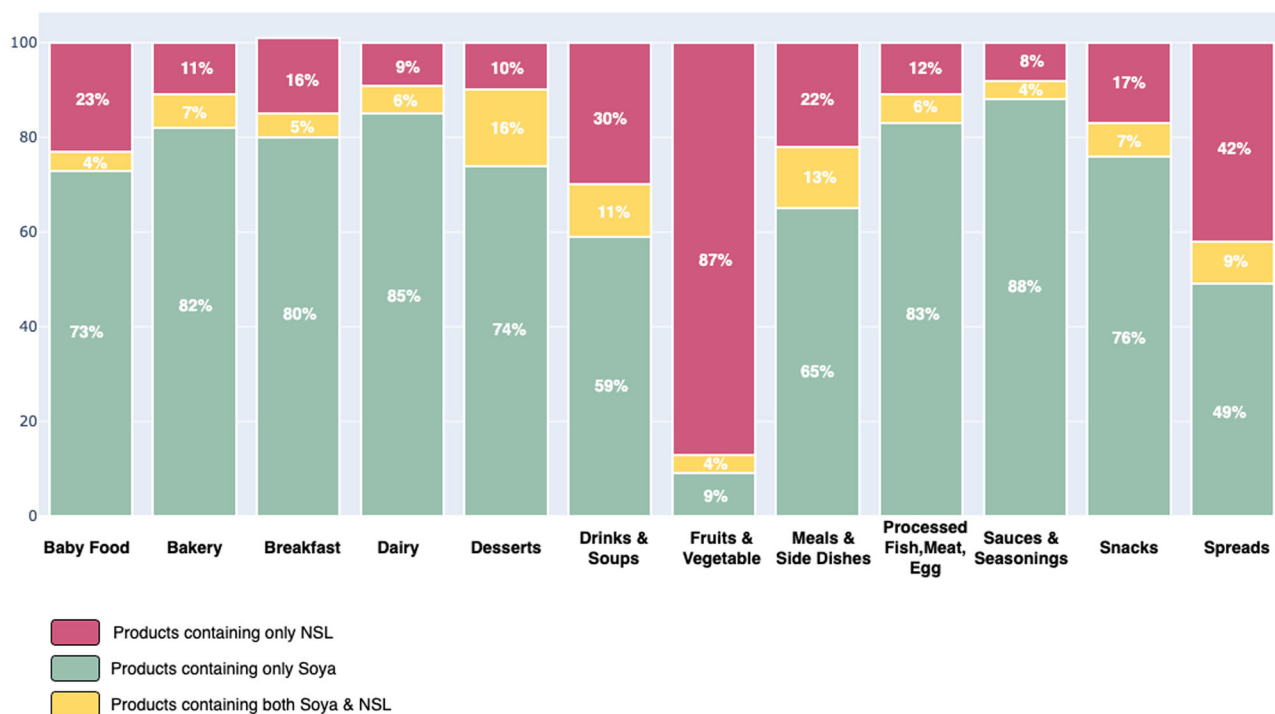


Fig. 1 | Shares of products containing soy-based ingredients or NSL-based ingredients in each market segment (%). The percentage sum may exceed 100% because some products have both soy- and NSL-based ingredients (see Fig. 3). The color intensity reflects the cumulative growth of product launches in the market

segment for soy and NSL ingredients over the decade. The market segment categories are those established in the Mintel-GNPD database and detailed in another work³².

several observations. Although soy dominates in terms of volume, the cumulative growth rate of products containing NSL ingredients is much greater than that of products containing soy, regardless of the market segment considered (see Table S2 in Supplementary Materials). Some products containing NSL ingredients experience a very high cumulative growth rate, particularly in the “dairy” segment, which is almost 12 times greater than that of soy-based products. The “desserts” segment is almost 9 times greater, and the “breakfast” segment is 7 times greater. More generally, these observations point to the growing interest among food companies in NSL ingredients³¹. Such growing interest, if confirmed over time, can favor a greater diversity of the legume species used.

Europe and southern Asia present a less unbalanced market between soy and NSL ingredients

The differences between products containing soy ingredients and those containing NSL ingredients are also more significant when we observe the share of these two categories in each main geographical area of our corpus (Fig. 2).

Soy ingredients are dominant in every geographical area compared to NSL ingredients. Nevertheless, some of these areas show a more balanced picture. This is particularly true for Europe and southern Asia, where approximately 40% of product launches contain NSL ingredients, showing a more balanced market between soy- and NSL-based products. Conversely, all other geographical areas show a much more asymmetrical market, largely in favor of soy-based products compared to NSL-based products. Thus, the share of products containing soy ingredients in the North and South American markets, as in Eastern and Southeastern American markets, is extremely dominant (approximately 80%). In the other geographical areas such as smaller and less well-documented markets in the Mintel database (which is notably the case for Africa), the share of packaged food products containing soy, although lower (over 60% and under 73%), remains dominant.

The interpretation of such differences across geographical areas is probably multifactorial in nature. The structuring of the different markets

may reflect differences in terms of food culture. For instance, soy products dominate Asian markets, except the southern Asian market, which includes India, a country where pulse (and particularly lentil) production and consumption levels are among the highest in the world^{36,37}. Moreover, such differences can also be interpreted as the consequences of different national or international public support schemes for pulse consumption, as is the case for Europe³⁸. Nevertheless, regarding Europe, this quasi-balance between soy and NSL food products can indicate a shift in the technological lock-in that European countries have encountered until now³¹⁻³³, which is beneficial for greater crop diversity in the processed food supply in the near future.

More generally, the overall structure of the corpus, whether in terms of market segments or geographical areas, reveals that a small number of species account for the bulk of legume-based packaged food innovations (i.e., new products or launches). This concentration stands in the way of greater diversity in processed food markets. The more we use a small number of species to produce a larger and growing variety of foods (soy is present in all market segments), the less room there is for the development of other species. This situation, partly resulting from historical and economic factors that lead to a lock-in situation, can undergo contemporary changes. However, to confirm an actual possible shift that would favor the use of NSL species in food offerings, we also need to look at the ways in which these species are used in product formulations.

Product-context use of legumes: entering the importance of the ingredient

To further assess the diversity of those legume species used in product launches, we examine the “product-context of use” of these species. In terms of diversity, we assume that it can be misleading to consider, at the same level, the food products that use these species for different reasons about which we do not know. Notably, from the point of view of product formulation, we may consider that the functional properties (technological, organoleptic, nutritional, etc.) derived from the parts of the species used account for more than the species itself. In this way, we propose to approach

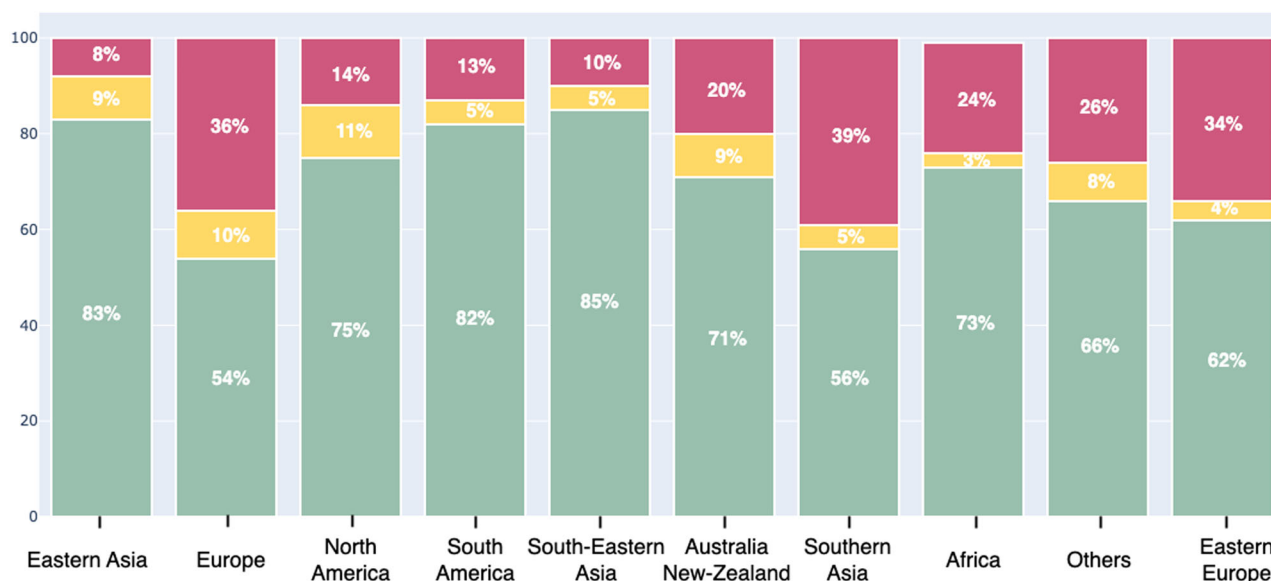


Fig. 2 | Soy- and NSL-based products in the main geographic areas covered by the corpus. The list of the countries covered in each geographic area is given in the Supplementary Materials (Table S4).

Table 2 | Soy-based and NSL-based ingredients mean position according to ingredient list length

Deciles	Nb. launches	NSL launches	NSL class share %	NSL ingr. mean position	Soy launches	Soy class share %	Soy ingr. mean position	Launches w. NSL and Soy
(1,7]	39,188	18,135	19.32	1.57	21,888	7.88	2.54	835
(8,11]	33,769	10,841	11.55	3.04	24,440	8.8	4.51	1512
(12,14]	30,859	8561	9.12	4.25	23,957	8.63	5.61	1659
(15,18]	43,771	11,170	11.9	5.4	35,385	12.74	6.81	2784
(19,21]	31,697	7558	8.05	6.5	26,432	9.52	8.21	2293
(22,24]	27,773	6278	6.69	7.56	23,705	8.54	9.34	2210
(25,29]	37,431	8246	8.78	9.46	32,678	11.77	10.64	3493
(30,35]	31,052	7117	7.58	11.79	27,343	9.85	12.13	3408
(36,46]	35,386	7772	8.28	15.13	32,057	11.55	13.45	4443
(47,278]	32,383	8206	8.74	25.6	29,771	10.72	17.74	5594

Here, “Deciles” column corresponds to the discretization into deciles of the ingredient list lengths of the products in the corpus. For each one of this class, we look at the number of products for which the length of the ingredient list corresponds to the decile (“Nb. Launches”), the average position of the soy or NSL ingredient in the list (“NSL/Soy ingr. mean position”), and the share these products represent of all products containing soy or NSL (“NSL/Soy Class Share”). The results are displayed for products containing both NSL-based and soy-based ingredients.

what we call the “product-context of use” by jointly analyzing the different positions of appearance of those species in ingredient lists.

A good starting point is the examination of where the identified species appear in ingredient lists. Regulations require ingredients to be listed in descending order of importance, with the first ingredient weighing the most and the last ingredient weighing the least. Hence, we assume that a species that is used only for a few of its functionalities (for instance, a treatment process aimed at extracting one or more of its parts, such as peptides, starches, and gelling compounds) is more likely to be found among the least important ingredients of an ingredient list (i.e., those weighing the least). This approach can be further refined by assessing whether or not the species identified in food products are part of the marketing pitch. We assume that the mention of the species on product packaging (in addition to its appearance in the ingredient list) gives higher specificity to the species used, as it is positively associated with the identity of the product. From this point of view, the differences between soy and NSL ingredients are quite striking.

Table 2 reports the mean position of soy- and NSL-based ingredients according to the ingredient list length, grouped in deciles. We observe that half of all the products (52%) containing NSL ingredients are concentrated

within the first four deciles; the first decile accounts for almost 20% of the products containing NSL ingredients. For products containing soy ingredients, this threshold is reached from the 6th decile upward.

More generally, soy-based ingredients tend to appear more frequently in food products with complex formulations (i.e., longer ingredient lists) and almost systematically at a higher rank (column “Soy ingr. mean position” in Table 2) than NSL-based ingredients. In all the deciles except for the last three, the mean position of NSL ingredients is always lower than that of soy ingredients. This finding means that NSL ingredients tend to appear more at the top of ingredient lists compared to soy ingredients, suggesting that the amount of the former used in the product formulations is probably greater than that of the latter used. This result can be explained by the fact that soy cracking has been much more widely studied than pulses in the field of food sciences and technology, particularly during the last decade³⁴. Research and development in this domain have led to a broader knowledge base for the various uses and functionalities of soy in comparison with other pulses/NSL. In view of this, our results can confirm that soy use is associated with a larger array of functional ingredients than NSL use. The likelihood of finding soy ingredients for use as additives in product formulations is likely

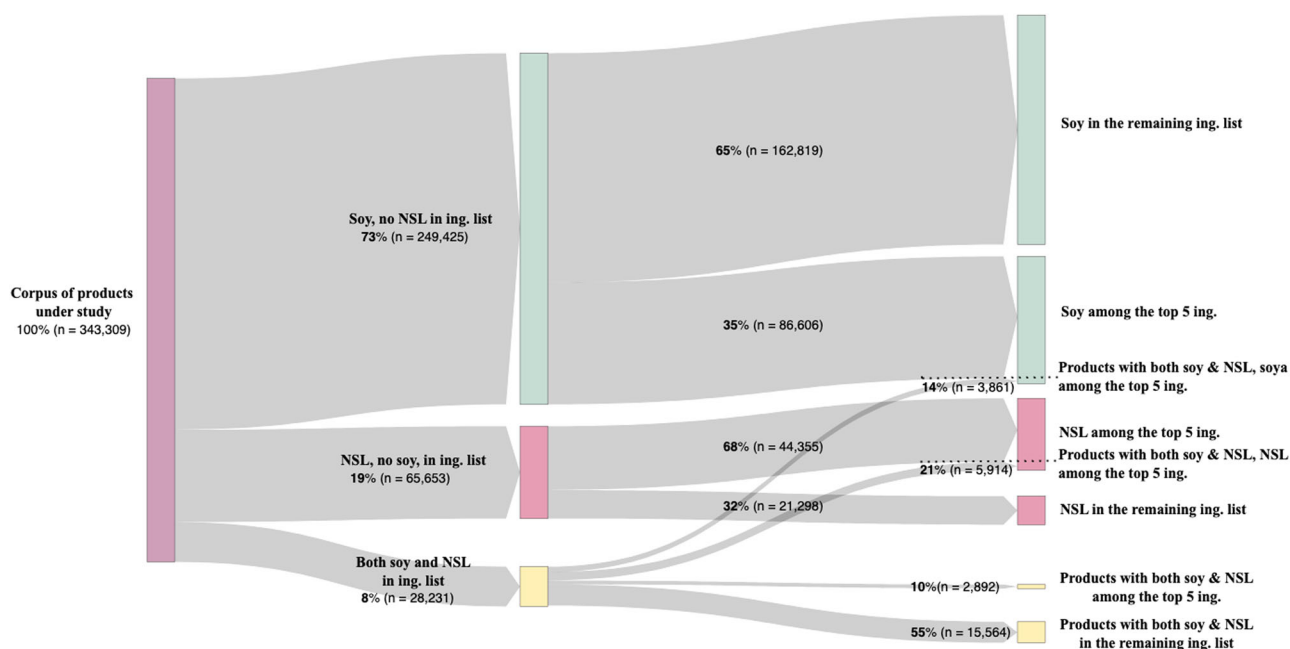
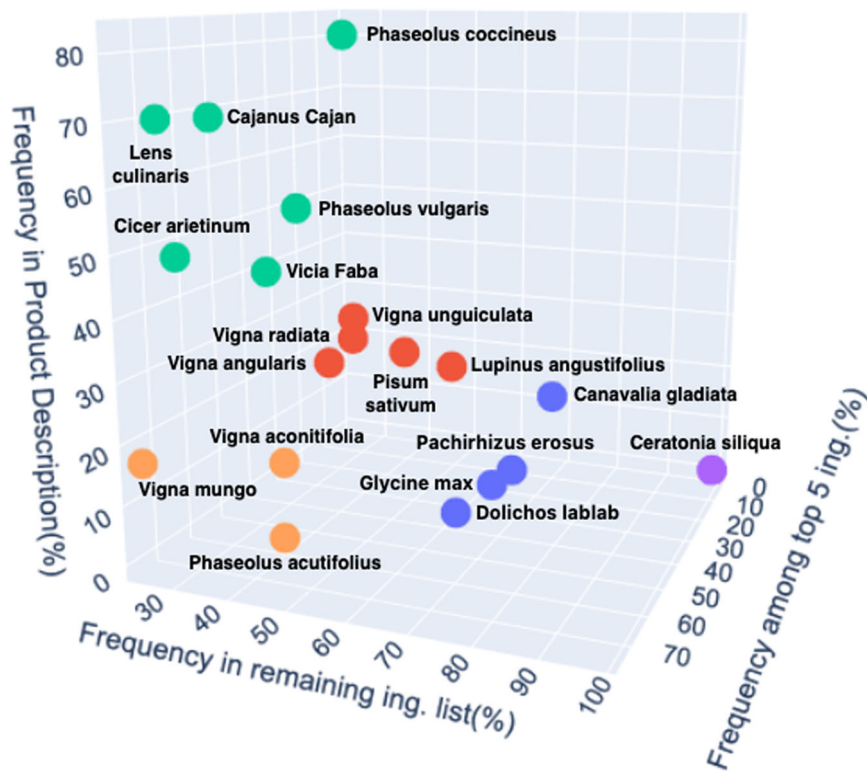


Fig. 3 | Sankey diagram of the dataset according to the position of soy or NSL ingredients (n = 343,344 food products). The dataset consists of the following 3 main subcorpora: products containing soy ingredients, products containing NSL ingredients, and products containing both types of ingredients. For each corpus, the

position of the ingredients in the list is provided—either among the first 5 first ingredients or among the remaining ingredients. For the small portion of products having both NSL and soy ingredients, their position at the top or in the remaining list can differ, resulting in four flows from this node.

Fig. 4 | 3D scatterplot of the most frequent legume species. Each species is plotted in a 3D graph according to its frequency of appearance in the first five ingredients (InFirst%), in the remaining ingredients (InRemList%), and in the product description (InDesc%). Each color represents a cluster resulting from *k*-means clustering (see the Methods section for the clusters requested according to results provided by the silhouette coefficients method⁴⁴).



greater than that for NSL ingredients, whose position is most often among the first ingredients on ingredient lists (Fig. 3).

Based on these new criteria, we refine our analysis by classifying species according to their frequency of appearance among the first (i.e., top) five ingredients. Figure 4 presents the results of this classification, which is

carried out using *k*-means clustering (see the Methods section) and interpreted by the food science experts involved in the analysis. The 5 resulting groups are identified by categorical colors, and species are displayed in a three-dimensional space showing their frequency of appearance (as a percentage) among the first 5 ingredients, the remaining ingredients, and on the

product packaging. In addition, to help interpret the results of this clustering method, we also examine the most frequent ingredient expressions associated with the species in each group.

In Fig. 4, a central axis structures the cluster distribution. This axis distinguishes the species mostly found among the first 5 ingredients and frequently in product descriptions (green cluster) from those found more frequently among the remaining ingredients and rarely cited in product descriptions (blue cluster). More precisely, at one end of this axis, we find a set of 6 species highlighted in green, namely, *Phaseolus coccineus* L., *Cajanus cajan* L., *Lens culinaris* Medik., *Phaseolus vulgaris* L., *Cicer arietinum* L., and *Vicia faba* L., characterized by a high frequency of appearance among the first ingredients and a high rate of mentions in product descriptions. These features lead us to suggest that product identity is more closely associated with those NSL species, regardless of their functional use, than with other NSL species. In that sense, species from this group can have a more positive impact on market biodiversity, as they are of key interest for the food industry, in comparison to species used only in terms of functional interest, and therefore can be substituted by other species. The most frequent ingredients associated with species from this group do not seem to indicate their fractional use. For example, in the case of *Lens culinaris* Medik., the most common ingredients mentioning this species are listed directly by its vernacular name, without mentioning specific parts (“lentils”, $n = 2522$; “red lentils”, $n = 1250$; and “green lentils”, $n = 818$). When this ingredient is associated with a processing term, the most frequent term is milling (“lentil flour”, $n = 1195$). The same is true for *Cicer arietinum* L. (“chickpeas”, $n = 7467$, and “chickpea flour”, $n = 3295$). This cluster gathers ingredients that seldom undergo processing.

At the opposite end of this central axis, plotted in blue, we find a group of 3 NSL species (*Canavalia gladiata* Jacq., *Pachyrhizus erosus* L., *Dolichos lablab* L.) and soy (*Glycine max* L.). They present the following opposite profile compared to that mentioned above: a low level of frequency among the first ingredients and in product descriptions. This cluster can also include *Ceratonia siliqua* L. (plotted in purple), which has been identified as a cluster in its own right due to its extreme behavior—it is hardly ever mentioned neither in product descriptions nor among the first ingredients. In this group, for the two most frequent species, *Glycine max* L. and *Ceratonia siliqua* L., the most frequently associated ingredients correspond to fractional uses (“soy oil”, $n = 58,584$; “soy lecithin”, $n = 50,490$; “soy protein”, $n = 18,569$; “locust bean gum”, $n = 4903$; and “carob bean gum”, $n = 2242$).

This axis, which contrasts species according to their frequency of appearance (within the first 5 ingredients and in the product description), may also tend to oppose different product contexts of use of species, thus bringing us back to our initial hypothesis: the more frequently that a species is used in a fractional way, the more likely it is to be found among ingredients of lesser importance (in terms of volume and therefore rank and level) in ingredient lists and the less prominence it is given in product packaging.

Hence, the case of the median cluster (plotted in red in Fig. 4) is very interesting. Here, we find species characterized by a balanced score between their frequency of appearance among the first and remaining ingredients but not systematically mentioned in product descriptions (*Pisum sativum* L., *Vigna unguiculata* L., *Vigna angularis* L., *Vigna radiata* L., and *Lupinus angustifolius* L.). According to our main hypothesis, this median position between species can reveal various strategies of the food industry for those species that can become more “identical” or for those “generic” species, according to the future uses that such species will encounter. In other words, most of the time, these species can be used as effective key components for product formulation. The analysis of the most frequent ingredients quoting the most major species of this group seems to substantiate this observation. For *Pisum sativum* L., the two most common ingredients are “peas” ($n = 10,132$) and “pea protein” ($n = 6423$), and for *Lupinus angustifolius* L., the most common ingredient is lupin flour ($n = 1095$).

Finally, a group of three species (plotted in orange in Fig. 4), made up of *Vigna aconitifolia* Jacq., *Vigna mungo* L., and *Phaseolus acutifolius* L., seems to be opposed to the first group described (plotted in green) due to the

weaker propensity of these species to be cited in product descriptions. These “discrete” species have a very low frequency of appearance, but the analysis of the most cited related ingredients brings them closer to the first group. For example, the most common ingredients referring to *Vigna mungo* L. mention the species by its vernacular name (“black gram lentils”, $n = 1412$, and “black lentils”, $n = 114$), and when a process is mentioned, in most cases, it concerns flour, the resulting product from grinding possibly coupled with sieving (“black lentil flour”, $n = 56$). We observe the same phenomenon for *Phaseolus acutifolius* L. (“tepariy bean flour”, $n = 128$, and “tepariy beans”, $n = 25$).

Discussion

Although there is abundant and growing scientific literature on the benefits of biodiversity for sustainable agrifood systems, this paper constitutes the first attempt to assess species diversity in the food market through the processed/packaged food products launched in the main regions of the world. We develop original text-mining analysis methods for tagging species and interpreting species diversity through the food ingredients used by the industry. We base this approach on the different positions of the species under study within ingredient lists and from additional descriptions on product packaging.

With a focus on legume species, a botanical family at the heart of sustainability issues in contemporary agrifood systems, this work compares the development of soy in processed food markets with the development of NSL. By analyzing the ingredient compositions of approximately 350,000 food products launched worldwide, we identify more than 30 different NSL species used in product formulation, which suggests consistent species diversity. Nevertheless, we mitigate this issue by taking an in-depth look at the frequency of appearance and ways in which such species are used in product formulation and then promoted on packaging. Through this first analysis, we can assess to what extent the industrial food market is concentrating on certain dominant species, which can hamper the development of species diversity. Indeed, one of our main hypotheses regarding path dependency is that the more a market locks in a certain species, the more difficult it becomes for other species to emerge, and this lock-in situation becomes a major hindrance to increased species diversity. We thereby confirm, at the food market level, what other studies have confirmed at the crop field level: there is still a strong lock-in situation around soy-based products, despite the perceptible shift in favor of other pulses. In this regard, our results particularly highlight the specificity of the European market, which presents a more balanced use of soy and NSL ingredients.

Furthermore, we assume that the position in ingredient lists of ingredients related to soy or NSL is a proxy of the ways in which these species are used. Considering this position (in the ingredient list) is a way to distinguish species that are more valued in their entirety as opposed to more “flexible” species used for their ability to be decomposed into various functional ingredients. By combining this information with the ways in which species are promoted or not promoted on packaging, we suggest that the notion of species and, therefore, of biodiversity, tends to be replaced by that of *functionality*. The food industry’s lack of interest in highlighting the species simply as it is, and before any processing, reflects an agrifood system where the production is increasingly valued by compositional/processed foods, a paradigm where food is first considered a technological process of assembling elements fulfilling a desired goal such as increasing protein level, improving texture, avoiding specific flavors, etc.¹⁶. If this work constitutes a first step toward a deeper analysis of what we call “the product-context use” of species in food products, then further research should undertake a deeper assessment of the processing profile of such ingredients.

This work also demonstrates the interest and feasibility of analyzing the agrifood market supply at the ingredient scale. While some other works studying the composition of food products have addressed food safety²¹ and climate issues³⁹, here, we focus on species diversity issues. Identifying how processed food impacts agrifood systems remains a key challenge for informing public policymakers and consumers. We argue that this type of analysis, at the ingredient level, can serve as a tool for public policy to steer

agrifood markets toward more sustainable goals. However, this situation depends on the availability of food databases and accessible controlled vocabularies and ontologies. Except for the USDA-branded food database and the crowd-sourced Open Food Facts database, the most extensive food databases required to conduct such an analysis are privately owned. Despite the intrinsic quality of these databases, whether for research purposes or to support public policy, this situation does not guarantee the control and transparency of information from collection to data processing. In this context, benefiting from controlled vocabularies validated by the academic community on the ingredients linked to processed foods can not only ensure transparency but also constitute a powerful tool for public policy⁴⁰. We also advocate for the need for enriched public databases on the processed food supply and for future works to confront this food supply with consumer purchases to analyze the market failure of species diversity. Furthermore, this work reveals several additional perspectives that we briefly discuss hereafter.

Firstly, identifying the species used by the food industry is of particular importance because agricultural diversity is a main lever for sustainability, and food outlets remain the main driver of cultivation choices by farmers. Beyond this aim of species diversity assessment, text-mining methods for ingredient lists (and any other information from packaging) allow us to follow the market development of any crop species, particularly those that are expected to be developed, such as pulses. Future work can further analyze the processing approach applied to crops to identify the ingredients and determine the types of processed food products under development. What a firm offers on the market reveals its know-how and reflects the technological paradigm that prevails in a given sector. Hence, this information can also be useful for science and innovation policies in the agrifood sector to support preferred technological paradigms that meet societal expectations regarding, for instance, less processed or minimally processed food with “cleaner” formulations. This work also highlights the need to develop a common vocabulary or ontology that favors transparency and contributes to debates about food classifications. For instance, a clear classification can allow for a better distinction between plant- or chemical-derived ingredients. The outcome of this work constitutes a case study that can be integrated into ontology development, such as TransformON or FoodON, which are interconnected databases in the agrifood system^{40,41}.

Secondly, a more complete picture requires further work. First, the Mintel database's coverage of countries is not complete. Second, we need to look at all species and consider other food supplies, including the indigenous supply, which particularly concerns less developed countries compared with the processed/packaged food supplies provided in most modern countries. However, the availability of these data is challenging. Concerning processed/packaged food, we highlight several impediments, such as the absence of global regulation on the ways in which to name most ingredients that complicate the analysis and call for the development of specific dictionaries and ontologies. This is also true for indigenous supplies, but in addition, qualitative inquiries are required because no database exists for more traditional food procurement. In addition, the issue of the importance of the ingredients in the product and, more largely, in the supply (each product also having various weights) remains in question. The position of an ingredient among other ingredients in a list is a way to inform on this “importance” but requires complex calculus to assess this biodiversity if we aim to consider all species. Such a complete picture requires the use of other metrics, such as the Herfindahl index or similar metrics.

Thirdly, one main hypothesis of our study is that the market incentivizes farmers to make certain cultivation choices. Therefore, rarely used species are rarely cultivated. This coevolution process creates a path dependence that favors the development of major crops. If globally assessing the species diversity used in the industry provides first-order information to confirm this path dependence, then looking at this issue within each country can inform policymakers regarding supporting new crops for more diversified agroecosystems around major species that may differ across countries. The issues related to the diversification of crop rotations differ across regions. However, to further explore this issue, we need to know the origin of

ingredients (at least for the first, more important, ingredients) to compare them with the geographical places where they are retailed. However, the mention of the geographical origin of the main ingredients is not mandatory. Geographical information about products usually concerns the headquarters of the retailing firm and not the origin of the ingredients themselves.

Finally, another challenge concerns how to confront the food supply provided by the industry considering the purchases of consumers and even the different diets of consumers (including catering and restaurants, for instance). This issue also requires further methodological work, as there is no easy solution to compare *launch* data with *purchase* data. This is likely why the literature is unclear regarding the failure rate of new products launched on the market⁴². When studies focus on consumer purchases in retail channels with, for instance, Kantar data, it is not known which products available in retail channels are not purchased, and thus, we do not know which new products from the food industry (and, notably, those products that can meet societal challenges) are not chosen by consumers. Those failed products then exit the market after several months and are thus not studied. However, it is interesting to understand the profile of those products and to investigate the reasons why they do not meet consumer demand. Price can be one reason for this difference, but probably not the only reason. To confront *purchase* and *launch* data, one solution is to merge Kantar and Mintel product data according to the same identifier, for instance, product barcodes. Nevertheless, this approach is currently not possible because the Kantar data do not include barcodes and, above all, because the regulation of barcodes allows the reuse of a barcode already used for a previous product that disappeared from the market, bringing about the increased risk of error. Therefore, another solution is to use text-mining methods to determine the correspondence between products at the product description level or to compare the ingredient lists. These text-mining methods also allow us to follow over time the changes in the ingredient list of a branded product that maintains a similar name or product description.

All these perspectives suggest that text-mining methods open a new research agenda for furthering the understanding of the evolution of food markets, particularly for following the evolution of crop species in foods.

Methods

Data sourcing

The data are retrieved from Mintel's Global New Products Database (GNPD), which tracks food product launches in more than 80 countries. The Mintel-GNPD database features approximately 7 million registered branded products (on average, 40,000 product descriptions have been added each month in recent years). The Mintel-GNPD database is currently the only database that provides such global coverage of packaged/processed food offerings, with detailed information at the product level⁴³. The map in the Supplementary Materials section provides an overview of the geographical coverage of the Mintel database due to its global shopper network (Fig. S1). With emergent and developing countries, Western countries have benefited from the large coverage of the Mintel database, which has progressively enlarged, mainly since 2010.

By comparison, the Open Food Facts database lists only approximately 2,800,000 branded products, and the United States Department of Agriculture (USDA) Global Branded Food Products database lists more than 3,500,000 products. However, these two databases are nurtured by volunteer declarations, mainly from consumers or industry firms, and involve a few countries. The Mintel-GNPD database relies on a census organized by shoppers hired in each country of their coverage to identify any new product launched on the market and to report the full information present on the product packaging. In this way, Mintel data provide a picture of the packaged food supply based on what products are provided by the food industry in markets worldwide. Our objective is to assess how the food industry is moving toward greater pulse diversity, and we choose to study food launches rather than food purchases. In addition, confronting the offer and demand data leads to unsolved methodological problems, which we discuss in the last section.

Raw Input:

```
'''tomato* (tomato pulp*, tomato concentrate*), water, rehydrated red kidney beans* (11%),
courgettes* (8%), carrots* (7%), onions* (6%), peppers* (6%), soy proteins* (5%), rehydrated
chickpeas*, cold extracted extra virgin olive oil*, spices*, sea salt, garlic*, basil*
*ingredients from organic farming'''
```

----- *** -----

Dictionary Output:

```
{'75_1': {'rawing': 'tomato', 'level': 0, 'comment': ['ingredients from organic farming']},
'75_2': {'rawing': 'tomato pulp', 'level': 1, 'comment': ['ingredients from organic
farming']}, '75_3': {'rawing': 'tomato concentrate', 'level': 1, 'comment': ['ingredients from
organic farming']}, '75_4': {'rawing': 'water', 'level': 0}, '75_5': {'rawing': 'rehydrated
red kidney beans', 'level': 0, 'prop': ['11%'], 'comment': ['ingredients from organic
farming']}, '75_6': {'rawing': 'courgettes', 'level': 0, 'prop': ['8%'], 'comment':
['ingredients from organic farming']}, '75_7': {'rawing': 'carrots', 'level': 0, 'prop':
['7%'], 'comment': ['ingredients from organic farming']}, '75_8': {'rawing': 'onions',
'level': 0, 'prop': ['6%'], 'comment': ['ingredients from organic farming']}, '75_9':
{'rawing': 'peppers', 'level': 0, 'prop': ['6%'], 'comment': ['ingredients from organic
farming']}, '75_10': {'rawing': 'soy proteins', 'level': 0, 'prop': ['5%'], 'comment':
['ingredients from organic farming']}, '75_11': {'rawing': 'rehydrated chickpeas', 'level': 0,
'comment': ['ingredients from organic farming']}, '75_12': {'rawing': 'cold extracted extra
virgin olive oil', 'level': 0, 'comment': ['ingredients from organic farming']}, '75_13':
{'rawing': 'spices', 'level': 0, 'comment': ['ingredients from organic farming']}, '75_14':
{'rawing': 'sea salt', 'level': 0}, '75_15': {'rawing': 'garlic', 'level': 0, 'comment':
['ingredients from organic farming']}, '75_16': {'rawing': 'basil', 'level': 0, 'comment':
['ingredients from organic farming']}}
```

Fig. 5 | Example of a parsed list of ingredients. Illustration of the output dictionaries obtained by the parser⁴³. Each ingredient has an identifier indicating its order of appearance. Information such as the initial name of the ingredient, the

depth to which it appears in the list of ingredients (the ingredient of an ingredient), the comment(s) associated with this ingredient, and any proportional information is retained.

Data are retrieved from the Mintel database via queries targeting legume-related products (legume, pulse, bean, pea, lentil, chickpea, etc.) in all food and drink product categories (except alcoholic beverages and mineral or source waters) and concerning launches from 2010 to 2021 (see screen capture of the query built in Mintel search engine provided in the Supplementary Materials, Fig. S2).

We keep approximately 350,000 products launched during the period 2010–2021 worldwide, with at least one pulse species or soy ingredient. The data are extracted in the form of a tabulated file gathering all the metadata available for each food product, such as the ingredient lists, production and marketing locations, types of product launches (new product, range extension, reformulation, repackaging or relaunch), targeted markets, and claims mentioned.

Data parsing

As ingredient lists are provided in the form of variable-length character strings, which are, by definition, difficult to manipulate and exploit by computational methods, we specifically design a parsing algorithm for this work. This algorithm, which is scripted in the Python language, is detailed in Salord et al.⁴³ and allows for the transformation of ingredient lists extracted from the Mintel database into structured and standardized ingredient dictionaries (in json format). Furthermore, this algorithm relies on a set of standardization rules based on the empirical observation of variations in the syntactic notations of these ingredient lists.

These normalization rules are translated within the algorithm in the form of regular expressions, thereby making the procedure inexpensive in

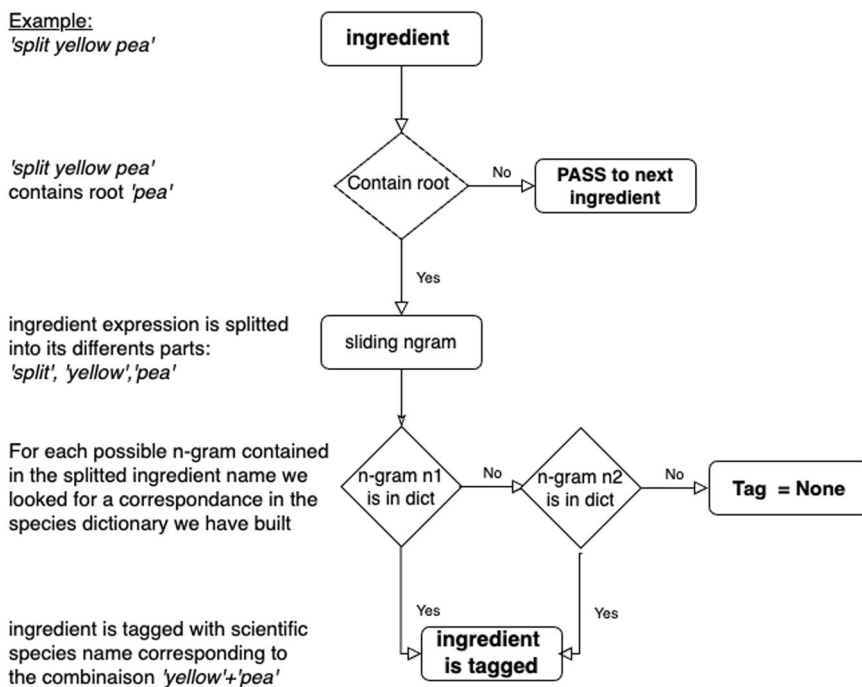
terms of computing resources. Furthermore, raw ingredient lists can be transformed into ingredient dictionaries without any loss of information. Hence, we keep key features such as the rank of the ingredient within the ingredient list (being the first, second, etc., ingredient in the ingredient list) and the depth of appearance (being or not being in brackets) as ingredients of another ingredient. A depth of 1 indicates that the ingredient is not mentioned as being an ingredient of another ingredient. The following insert, Fig. 5, illustrates such considerations.

For our analysis, we consider legume-based ingredients of level-1 depth, and we distinguish those among the first five ingredients (we called the “In Top List”) from the remaining ingredients (we called the “Remaining list”). This distinction is adopted considering both the average distribution of soy-based ingredients and the ingredients related to other legume species within ingredient lists, which have variable lengths.

Ingredient tagging

After structuring the unstructured textual data extracted from the ingredient lists, we are able to identify only the legume species present in the corpus. To perform this tagging stage, we use a mixed method relying both on text-mining techniques and on the scientific expertise of a group of six food scientists who contributed to designing a dictionary of tagging rules. One of the main difficulties in identifying mentions of plant species (in our particular case, soy and non-soy legume (NSL) species) in ingredient lists relates to the fact that these species are almost systematically designated by their vernacular name or by part of their vernacular names, with the latter sometimes being misspelled. In addition, some of these names can be

Fig. 6 | Flow chart of the tagging process. Here, each stage of the tagging process is illustrated. The dictionary mentioned in the chart is given as supplementary material (Table S3).



sources of error or confusion depending on the context in which they are cited. To overcome such issues and rigorously associate ingredients with the scientific name of the plant species to which they refer, the tagging method used is based on the identification of those stems most often associated with a given species and on the context in which they are written. Thus, we first extract from the correctly parsed ingredient lists all ingredients containing stems that can be associated with a pulse species (“pea”, “bean”, “lens” or “lentil”, “gram”, etc.) as well as the words immediately preceding and following them. Such words are of utmost importance, as they tell us about the type of species mentioned. For example, for “black-eyed peas” and “grass peas”, it is erroneous to deduct the presence of *Pisum sativum* L. simply from the presence of the stem pea. Here, each of the terms preceding “peas” refers to a different pulse species—*Vigna unguiculata* L. for “black-eyed peas” and *Lathyrus sativus* L. for “grass peas”.

This fine-grained analysis is conducted by pulse food science experts, who helped identify, for each combination of a given stem and its preceding and following words, the correct scientific name of the related species. When no consensus is reached among the experts or when it is impossible to clearly identify a plant species, we use an external gold standard database, such as the Catalog of Life, or we include outside experts. This is particularly the case when it comes to understanding and correctly identifying the rich diversity of pulse species involved in “dahl”, a traditional Indian meal. If, nevertheless, no consensus is reached or the identification of a species is not assured, then the ingredient is considered ambiguous.

Nevertheless, some other ambiguities remain between fresh and dried legumes. Firms sometimes do not distinguish between the use of fresh and dried legumes, although the term “pulses” is reserved for pulses harvested once their grains have dried. To avoid any ambiguity, because it is sometimes impossible to deduce from food product compositions whether the legumes used are fresh or dry, we label the pulse species identified in our corpus as NSL. This important annotation work results in the production of a dictionary of expressions, which we provide as supplementary data (Table S3). This dictionary is then translated into a computer automaton to label all the ingredients presenting the same form in our corpus that refers to a legume species. The following Fig. 6 summarizes this tagging procedure.

For the purpose of this study, the same tagging process is applied to product descriptions to identify the legume species mentioned on packaging beyond the ingredient list. Such mentions tell us how a particular species is

promoted and how it is used to add value to the product. All food product descriptions (as reported in the Mintel database) are tokenized and cleaned before applying the same set of labeling rules that we use for ingredient lists. For this specific stage, no expert supervision is mobilized.

Final dataset

At the end of these processing stages, once ambiguous mentions of species are discarded, we obtain a dataset of 343,309 products, not considering the following specific cases ($n = 4800$): (i) ingredients that refer ambiguously to a legume species (as in the expression “pulse bean”), (ii) false friends (as in the expression “coffee bean” or “coco pea”), (iii) legume species that do not enter into the subfamily of pulses such as *Arachis hypogaea* L., and (iv) cases where the tagger returned an error, i.e., a “none” value.

Our final dataset is composed of three subcorpora: (i) products whose lists of ingredients contain soy-based ingredients and no NSL-based ingredients ($n = 249,425$), (ii) products whose lists of ingredients contain NSL-based ingredients and no soy-based ingredients ($n = 65,653$), and (iii) products containing both soy-based and NSL-based ingredients ($n = 28,231$).

Statistics. The descriptive statistics presented in the paper are computed considering these three subcorpora at the following different scales: (i) at the product level, identifying the main characteristics of products containing legumes (soy and NSL), and (ii) at the species level, characterizing the way in which they are used in packaged food product launches. This second level constitutes the first step toward a more systematic analysis of the uses of these plant species. The present study involves an initial effort to categorize these species according to their frequency of appearance in the food industry (whether they are frequent, infrequent, or rare in food product formulations) and their ranking (whether they appear among the first or last items on ingredient lists based on the assumption that ingredient lists are ordered according to the weight of the ingredient). This categorization work is carried out by performing k-means clustering on the frequencies of appearance of legume species among the first ingredients on ingredient lists and among the last ingredients on ingredient lists, provided that these species appear at least one hundred times in the corpus (40 randomly chosen centroids for 3000 iterations, w a maximum of 2 to 11 clusters asked). The selection of the number of

clusters to be retained is made using the silhouette coefficient method ($S(i) = \frac{b(i)-a(i)}{\max\{a(i)-b(i)\}}$), the graphical output of which is provided in the Supplementary Materials (Fig. S3 and Table S5). The results of this work can be found in Fig. 4 in the “Results and Discussion” section.

Data availability

The list of the product identifiers (used by MINTEL) and the corresponding species identified is free and available on this data verse: <https://doi.org/10.57745/KKAY1I>.

Code availability

The algorithm used for parsing the ingredient lists of the food products retrieved from the MINTEL database is freely available at <https://github.com/Pythrix/FOODCOP.git> and described in Salord et al.⁴³

Received: 30 November 2023; Accepted: 30 August 2024;

Published online: 28 September 2024

References

- Rockström, J., Edenhofer, O., Gaertner, J. & DeClerck, F. Planet-proofing the global food system. *Nat. Food* **1**, 3–5 (2020).
- Campbell, B. M. et al. Agriculture production as a major driver of the Earth system exceeding planetary boundaries. *Ecol. Soc.* **22**, art8 (2017).
- Tilman, D. & Clark, M. Global diets link environmental sustainability and human health. *Nature* **515**, 518–522 (2014).
- Benton, T. G., Bieg, C., Harwatt, H., Pudasaini, R. & Wellesley, L. Food System Impacts on Biodiversity Loss. *Chatham House - The Royal Institute of International Affairs Chatham House* (2021).
- Renard, D. & Tilman, D. National food production stabilized by crop diversity. *Nature* **571**, 257–260 (2019).
- Pilling, D. & Bélanger, J. *The State of the World's Biodiversity for Food and Agriculture* (FAO, 2019).
- Thomine, E., Mumford, J., Rusch, A. & Desneux, N. Using crop diversity to lower pesticide use: socio-ecological approaches. *Sci. Total Environ.* **804**, 150156 (2022).
- Bedoussac, L. et al. Ecological principles underlying the increase of productivity achieved by cereal-grain legume intercrops in organic farming. A review. *Agron. Sustain. Dev.* **35**, 911–935 (2015).
- Zimmerer, K. S. & De Haan, S. Agrobiodiversity and a sustainable food future. *Nat. Plants* **3**, 17047 (2017).
- Gil, J. Forgotten crops confer resilience under climate change. *Nat. Food* **4**, 275–275 (2023).
- Labeyrie, V. et al. The role of crop diversity in climate change adaptation: insights from local observations to inform decision making in agriculture. *Curr. Opin. Environ. Sustain.* **51**, 15–23 (2021).
- Mustafa, M. A., Mabhaudhi, T. & Massawe, F. Building a resilient and sustainable food system in a changing world—A case for climate-smart and nutrient dense crops. *Glob. Food Secur.* **28**, 100477 (2021).
- Altieri, M. & Nicholls, C. *Biodiversity and Pest Management in Agroecosystems, Second Edition*. (CRC Press, Boca Raton, FL, 2004).
- Beillouin, D., Ben-Ari, T., Malézieux, E., Seufert, V. & Makowski, D. Positive but variable effects of crop diversification on biodiversity and ecosystem services. *Glob. Change Biol.* **27**, 4697–4710 (2021).
- Webb, P. et al. The urgency of food system transformation is now irrefutable. *Nat. Food* **1**, 584–585 (2020).
- Leite, F. H. M. et al. Ultra-processed foods should be central to global food systems dialogue and action on biodiversity. *BMJ Glob. Health* **7**, e008269 (2022).
- Burlingame, B. & Dernini, S. *Sustainable Diets and Biodiversity: Directions and Solutions for Policy, Research and Action: Proceedings of the International Scientific Symposium Biodiversity and Sustainable Diets United Against Hunger, 3–5 November 2010, FAO Headquarters, Rome*. (FAO, Rome, 2012).
- Tilman, D. Extinction, climate change and the ecology of *Homo sapiens*. *J. Ecol.* **110**, 744–750 (2022).
- Köhler, J. et al. An agenda for sustainability transitions research: State of the art and future directions. *Environ. Innov. Soc. Transit.* **31**, 1–32 (2019).
- Campbell, B. M., Thornton, P. K. & Nelson, G. C. Upping our ambition for food system adaptation. *Nat. Food* **3**, 970–971 (2022).
- Ahuja, J. K. C. et al. InGIID: A framework for parsing and systematic reporting of ingredients used in commercially packaged foods. *J. Food Compos. Anal.* **100**, 103920 (2021).
- Sadler, C. R. et al. Processed food classification: conceptualisation and challenges. *Trends Food Sci. Technol.* **112**, 149–162 (2021).
- Rogers, E. M. *Diffusion of Innovations*. (Free Press, Collier Macmillan, 1983).
- Geels, F. W. Socio-technical transitions to sustainability: a review of criticisms and elaborations of the multi-level perspective. *Curr. Opin. Environ. Sustain.* **39**, 187–201 (2019).
- Raven, R., Bosch, S. Vden & Weterings, R. Transitions and strategic niche management: towards a competence kit for practitioners. *IJTM* **51**, 57 (2010).
- Toledo, Á. & Burlingame, B. Biodiversity and nutrition: A common path toward global food security and sustainable development. *J. Food Compos. Anal.* **19**, 477–483 (2006).
- Holden, N. M., White, E. P., Lange, M. C. & Oldfield, T. L. Review of the sustainability of food systems and transition using the Internet of Food. *npj Sci. Food* **2**, 18 (2018).
- Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K. & Bloem, M. W. Legumes as a sustainable source of protein in human diets. *Glob. Food Security* **28**, 100520 (2021).
- Ditzler, L. et al. Current research on the ecosystem service potential of legume inclusive cropping systems in Europe. A review. *Agron. Sustain. Dev.* **41**, 26 (2021).
- Foyer, C. H. et al. Neglecting legumes has compromised human health and sustainable food production. *Nat. Plants* **2**, 16112 (2016).
- Cusworth, G., Garnett, T. & Lorimer, J. Agroecological break out: legumes, crop diversification and the regenerative futures of UK agriculture. *J. Rural Stud.* **88**, 126–137 (2021).
- Magrini, M.-B. et al. Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agrifood system. *Ecol. Econ.* **126**, 152–162 (2016).
- Magrini, M.-B. et al. Pulses for sustainability: breaking agriculture and food sectors out of lock-in. *Front. Sustain. Food Syst.* **2**, 64 (2018).
- Magrini, M.-B. et al. Peer-reviewed literature on grain legume species in the WoS (1980–2018): a comparative analysis of soybean and pulses. *Sustainability* **11**, 6833 (2019).
- Magrini, M. B., Salord, T. & Cabanac, G. The unbalanced development among legume species regarding sustainable and healthy agrifood systems in North-America and Europe: focus on food product innovations. *Food Sec.* **15**, 187–200 (2023).
- Joshi, P. K. & Rao, P. P. Global pulses scenario: status and outlook. *Glob. pulse Scenar. Ann. N. Y. Acad. Sci.* **1392**, 6–17 (2017).
- Gurusamy, S., Vidhya, C. S., Khasherao, B. Y. & Shanmugam, A. Pulses for health and their varied ways of processing and consumption in India—a review. *Appl. Food Res.* **2**, 100171 (2022).
- Szczybyło, A., Halicka, E., Jackowska, M. & Rejman, K. Analysis of the global pulses market and programs encouraging consumption of this food. *Prob. World Agri.* **19**, 85–96 (2019).
- Clark, M. et al. Estimating the environmental impacts of 57,000 food products. *Proc. Natl. Acad. Sci. USA* **119**, e2120584119 (2022).
- Weber, M. et al. PO2/TransformON, an ontology for data integration on food, feed, bioproducts and biowaste engineering. *npj Sci. Food* **7**, 47 (2023).
- Dooley, D. M. et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Sci. Food* **2**, 23 (2018).

42. Rutkowski, I. P. Success and failure rates of new food and non-food products introduced on the market. *J. Market. Consum. Behav. Emerg. Market.* 52–61 <https://doi.org/10.7172/2449-6634.jmcbem.2022.1.4> (2022).
43. Salord, T., Magrini, M.-B. & Cabanac, G. Packaged foods with pulse ingredients in Europe: a dataset of text-mined product formulations. *Data Brief.* **42**, 108173 (2022).
44. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

Acknowledgements

The authors thank the anonymous referees for their helpful comments, which improved the quality of the paper. This research was supported by the PPR SPECIFICS project (ANR-20-PCPA-0008) funded by the “Growing and Protecting Crops Differently” French Priority Research Program (PPR-CPA), which is part of the national investment plan operated by the French National Research Agency (ANR) and the Occitanie Region (KING project). We thank Alice Thomson-Thibault for her English editing of the first version of the manuscript.

Author contributions

T.S., M.B.M.: conceptualization, methodology, investigation, formal analysis, writing original draft, reviewing, and editing. M.B.M.: supervision, funding acquisition. T.S., G.C.: investigation, code writing, data visualization. V.L.P., M.J.A., C.B., A.B., V.M.: investigation, formal analysis, writing—reviewing. M.W.: formal analysis, writing—reviewing. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41538-024-00305-7>.

Correspondence and requests for materials should be addressed to Tristan Salord or Marie-Benoît Magrini.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024